Analysis of data from multimodal chemical characterizations of plant tissues

Dissertation

zur Erlangung des akademischen Grades

doctor rerum naturalium

(Dr. rer. nat.)

im Fach Chemie

Spezialisierung Physikalische und Theoretische Chemie

eingereicht an der

Mathematisch-Naturwissenschaftlichen Fakultät der

Humboldt-Universität zu Berlin

von

M. Sc. Sabrina Maria Diehn

Präsidentin der Humboldt-Universität zu Berlin

Prof. Dr.-Ing. Dr. Sabine Kunst

Dekan der Mathematisch-Naturwissenschaftlichen Fakultät

Prof. Dr. Elmar Kulke

Gutachter/in:
1. Prof. Dr. Janina Kneipp
2. Prof. Dr. Ulrich Panne
3. Prof Dr. Stephan Seifert

Tag der mündlichen Prüfung: 07.12.2020

Abstract

The pre-processing and analysis of spectrometric and spectroscopic data of plant tissue are important in a wide variety of research areas, such as plant biology, agricultural science, and climate research. The focus of this thesis is the optimized utilization of data from plant tissues, which include matrix-assisted laser desorption/ionization mass spectrometry (MALDI-TOF MS), Raman spectroscopy, and FTIR spectroscopy. The ability to attain a classification using these methods is compared, in particular after combination of the data with each other and with additional chemical and biological information. The discussed are concerned with the investigation and classification within a particular plant species, such as the distinction of samples from different populations, growth conditions, or tissue substructures. The samples comprise grass pollens from large greenhouse experiments and from environmental samples, as well as tissue sections of the species *Sorghum bicolor* and *Cucumis sativus* from dedicated physiological and ecological studies. In this way, several results of this work contribute directly to these projects. The data were analyzed by exploratory tools such as principal component analysis and hierarchical cluster analysis, as well as by predictive tools that included partial least square-discriminant analysis and machine learning approaches.
Specifically, the results show that combination of the methods with additional plant-related information in a consensus principal component analysis leads to a comprehensive characterization of the samples. This was indicated by the discrimination of pollen from the grass species *Poa alpina* regarding different populations and environmental conditions. As another application of a multimodal analysis for classification, the combination of FTIR microspectra from individual pollen grains with their Raman mapping data and with MALDI-MS data of pollen extract is discussed. Moreover, as shown for Raman mapping data of tissue sections of *Sorghum bicolor*, the data of many Raman maps can be combined with other phenotypical data for an extensive insight into tissue biochemical composition.
Moreover, some important problems of the individual methods with regard to non-relevant variances in the respective data sets are addressed. In the case of Raman microspectra, a high non-Raman based variance in the data set can be caused by a strong fluorescence background, as is discussed here for the spectra of individual pollen grains. On the other hand, FTIR microspectra of individual pollen grains display scattering artifacts, unless they are obtained from samples that are embedded in paraffin, which, in turn, leads to paraffin signals in the pollen spectra. In MALDI-MS imaging data, the extracts of pollen mixtures can overlap and lead to superposition of unknown origin and extent. Optimized data pre-treatment strategies are discussed for each of these problems. They include advanced baseline correction methods by asymmetric least squares, decomposition by non-negative matrix factorization followed by spectra reconstruction, and normalization by extended multiplicative signal correction.

An important aspect is also the utilization of spectral maps for classification, as discussed here

for Raman and MALDI-MS imaging data. Possible approaches to a targeted data extraction from mapping data sets that enable classification on large amounts of mapping data are presented. Thereby, variance in the data sets can be reduced, and hierarchical classification can be refined. Using a large amount of Raman mapping data for a multivariate analysis, the possibility of recombining the classification outcome with the original mapping information is demonstrated. The results in this work indicate the relevance of the targeted utilization of spectrometric and spectroscopic data, and could be applied not only to plant-related topics but also to other analytical classification problems.

Zusammenfassung

Die Vorbehandlung und Analyse von spektrometrischen und spektroskopischen Daten von Pflanzengewebe ist in einer Vielzahl von Forschungsbereichen wie Pflanzenbiologie, Agrarwissenschaften und Klimaforschung wichtig. Der Schwerpunkt dieser Arbeit liegt auf der optimierten Nutzbarmachung solcher Daten, die Experimenten mit Matrix-assistierter Laser Desorption/Ionisierung Flugzeit-Massenspektrometrie (MALDI-TOF MS), Raman-Spektroskopie und FTIR-Spektroskopie unterschiedlichen Proben gewonnen wurden. Die einzelnen Methoden werden hinsichtlich ihrer Eignung zur Klassifizierung der Daten in spezifischen Fragestellungen miteinander verglichen. Darüber hinaus wird die Kombination der unterschiedlichen, komplementären Daten untereinander und mit weiteren chemischen und biologischen Informationen für die spektrenbasierte Unterscheidung sehr ähnlicher Proben eingesetzt. Die hier diskutierten Beispiele befassen sich mit Klassifizierung innerhalb einer bestimmten Pflanzenart, wie beispielsweise der Unterscheidung von Daten aus verschiedenen Populationen, Wachstumsbedingungen oder Gewebesubstrukturen. Die Proben umfassen Gräserpollen aus großen Gewächshausversuchen und aus dem Feld sowie Gewebeschnitte der Arten *Sorghum bicolor* und *Cucumis sativus* aus speziellen physiologischen und ökologischen Studien. Mehrere Ergebnisse dieser Arbeit tragen direkt zu diesen Projekten bei. Die Daten wurden mit explorativen Werkzeugen wie der Hauptkomponentenanalyse und der hierarchischen Clusteranalyse sowie mit Klassifizierungswerkzeugen analysiert, welche die Regression der partiellen kleinsten Quadrate (engl. *partial least squares-discriminant analysis*, PLS-DA) und Ansätze des maschinellen Lernens umfassen.
Die Ergebnisse zeigen, dass die Kombination der Spektren mit zusätzlichen Informationen in einer Consensus-Hauptkomponentenanalyse zu einer umfassenden Charakterisierung der Proben führt. Dies wird anhand der Unterscheidung von Pollen der Grasart *Poa alpina* hinsichtlich ihrer Zugehörigkeit zu verschiedenen Populationen und Umweltbedingungen belegt. Als weitere Anwendung einer multimodalen Analyse wird die Kombination von FTIR-Mikrospektren einzelner Pollenkörner mit bildgebenden Raman-Daten und mit MALDI-MS-Daten von Pollenextrakten diskutiert. Darüber hinaus können, wie hier für Raman-Karten von Gewebeschnitten von Sorghum bicolor gezeigt, die Daten vieler Raman-Kartierungen mit anderen phänotypischen Daten kombiniert werden, um einen umfassenden Einblick in die biochemische Zusammensetzung des Gewebes zu erhalten. Als ein weiterer Aspekt werden einige wichtige Probleme der einzelnen Methoden hinsichtlich nicht für die Unterscheidung relevanter Abweichungen brsprochen. Im Fall von Raman-Mikrospektren kann eine große Varianz im Datensatz durch einen starken Fluoreszenzhintergrund verursacht werden, wie hier für die Spektren einzelner Pollenkörner gezeigt wird. FTIR-Mikrospektren zeigen hingegen Streuartefakte, sofern einzelne Pollenkörner nicht in Paraffin eingebettet werden, was wiederum zu Paraffinsignalen in den Pollenspektren führt. In MALDI-MS-Bildgebungsdaten können sich die Extrakte von Pollenmischungen überlappen und zu einer überlagerung von

Peaks unbekannten Ursprungs führen, deren Ausmaß variiert. Für jedes dieser Probleme werden optimierte Datenvorbehandlungsstrategien vorgeschlagen und angewendet. Sie umfassen ausgewählte Basislinienkorrekturverfahren durch *asymmetric least squares*, durch nichtnegative Matrixfaktorisierung (engl. *non-negative matrix fatorization*, NMF) sowie Normierung durch extended multiplicative signal correction.
Für viele Anwendungen ist die Nutzbarmachung von spektralen Kartierungen zur Klassifizierung essentiell, wie hier für Raman- und MALDI-MS-Bilddaten diskutiert wird. Mögliche Ansätze für eine gezielte Datenextraktion aus Mapping-Datensätzen, die eine Klassifizierung großer Mengen von ortsaufgelösten Spektren ermöglichen, werden vorgestellt. Mit ihnen können die Varianz in den Datensätzen verringert und die hierarchische Klassifizierung verfeinert werden. Unter Verwendung einer großen Anzahl von Raman-Bilddaten für eine multivariate Analyse wird die Möglichkeit demonstriert, das Klassifizierungsergebnis mit der Ortsinformation jedes einzelnen Spektrums in vielen Karten zu rekombinieren. Die Ergebnisse dieser Arbeit zeigen die Relevanz einer gezielten Ausnutzung und Kombination mehrerer verschiedener spektrometrischer und spektroskopischer Daten für eine umfassende chemische Charakterisierung von Pflanzenproben. Viele der hier beschriebenen Vorgehensweisen können auch auf andere analytische Klassifizierungsprobleme angewendet werden.

Acknowledgments

Many people supported me during the last years and also many collaborators contributed directly or indirectly to this work. Here I would like to thank them for their support.
My special thanks go to Prof. Dr. Janina Kneipp, who supported me for several years since I joined the group during my bachelor studies. Even after I finished my Bachelor thesis, it feels that I never left until I joined the group again for my Master thesis, and subsequently as a PhD student. I am particularly grateful for this very interesting research topic. Chemometrics is not yet a usual topic in the studies of chemistry and I am very happy to had the freedom and time to learn a lot of skills that were useful in the PhD studies and will be in my future life.
I like to thank Prof. Dr. Stephan Seifert, who has been my mentor in my Bachelor and Master studies and introduced me to multivariate statistics and pollen analysis. Without the support, it would have been a much harder start for me in my PhD studies.
I thank Dr. Boris Zimmermann from the Norwegian University of Life Sciences (NMBU) for all discussions and the fruitful collaboration. I feel very grateful to be a part of a big collaboration network concerning the pollen analysis. In this regard, I would like to thank Prof. Dr. Achim Kohler, Dr. Murat Bağcıuğlu , Dr. Valeria Tafintseva, Prof. Dr. Siri Fjelheim, and Prof. Mikael Ohlson from NMBU for providing me with pollen samples, data, and algorithms that I used for the thesis. This collaboration increased my knowledge greatly about several aspects of pollen analysis.
Several experiments were carried out at BAM. Therefore, I would like to thank Prof. Dr Ulrich Panne and PD. Dr. Franziska Emmerling for the opportunity to perform the measurements in these facilities. I further thank Dr. Steffen Weidner and Dr. Franziska Lauer for the help with MALDI measurements and for the fruitful and smooth collaboration. I also thank Dr. Thomas Schmied for the support regarding the Raman experiment on a sometimes unpredictable device at BAM.
I thank Prof. Dr. Rivka Elbaum and her group at the Hebrew University for all the valuable contributions to several projects discussed in the thesis. Specifically, I would like to thank Fikadu Biru and Nerya Zexer for providing me with pollen samples and data. I want to acknowledge also all collaboration meetings with Prof. Dr. Rivka Elbaum and Nerya Zexer that give me the necessary insights into the biological facts of these projects.
I like to thank Dr. Ulrich Schade and Dr. Ljiljana Puskar for all the support for the FTIR experiments at BESSY II in Berlin. I further thank all former and present members of the Kneipp group. Particularly, I thank Victor Rodriguez Zancajo and Tom Lindtner for the nice collaborations and support that made especially the BESSY measurements a little bit less painful. I further like to thank Niclas Schauer and Simon Schröder for their collaboration in the investigation of the pollen samples. I thank Dr. Zsuzsanna Heiner and Ingrid Liedkte for providing me with data and fruitful discussion that helped me optimizing the data analysis. I like to thank Annette Rothe, Barbara Franke, and Stefanie Sellons for their administrative and

technical assistance.
I am very grateful for my colleagues to accompanied me during the last years at the university, but also outside the workplace. I like to thank Dr. Fani Madzharova, Dr. Vesna Zivanovic, Dr. Cecilia Spedalieri, Dr. Gergö Szekeres, Tom Lindtner, Victor Rodriguez Zancajo, Dr. Marcin Rybicki, Dr. Daria Galimberti, Stephen Leach, Christopher Sheldon, Nicole Mancini, Marcel John, Dr. Maristella Alessio, Dr. Jakob Kottmann, and Dr. Arpan Kundu. Also, during the writing of the thesis and the Corona-Lockdown the regular Zoom "Lunch meetings" helped me to stay sane. I want to thank again Fani Madzharova, Vesna Zivanovic, Cecilia Spedalieri, Gergö Szekeres, Tom Lindtner, Victor Rodriguez Zancajo, Daria Galimberti, Christopher Sheldon, and Tassilo Waniek for proofreading chapters of this dissertation.
I particularly like to thank my friends who helped me through the last years and specifically, my grandparents. Without them, it would not have been possible to get through my studies.

Contents

1 Motivation and structure of the thesis

In the last decades, spectrometric and spectroscopic methods have become important tools for the investigation of plant samples of various origins.[1–7] As an example, a whole community of researchers has founded the International Society for Plant Spectroscopy. A biannual conference on plant spectroscopy has been taking place since 2017.[8] This illustrates the relevance of utilization and evaluation of spectroscopic plant data.
Particularly, the accurate classification and characterization of pollen regarding their chemistry not only at the species level,[3–5,9–13] but also on the level of sub-species variance using spectroscopic methods are of great importance.[14–18] The molecular understanding of the pollen chemistry and their differences according to subspecies variation can, e.g., help to optimize crops and is of great importance for agricultural science, plant science, or climate studies. The exact composition of the pollen, including the unique biopolymer sporopollenin coat are not fully elucidated and most likely species-specific.[19] UV/Vis,[20,21] FTIR,[22] Surface-enhanced Raman scattering (SERS),[19] and nuclear magnetic resonance (NMR)[23] have contributed to the characterization of sporopollenin and the other pollen consituents. The interior of a pollen grain contains the germ cells, starch, and liquid bodies as well as typical cell compartments, such as the endoplasmic reticulum, vacuoles, and mitochondria.[24,25] The species identification is limited if pollen grains have similar size and shape. For example, the Poaceae family contains several thousand plant species,[26,27] where species identification is challenging due to their similar size and shape. Variation in pollen quality due to environmental influences experienced by the parental plants within the same population is usually evaluated by viability tests and assessment of germination rates.[28–30]
Here, the variances within pollen from closely related species and within the same species will be assessed using mass spectrometry and different vibrational spectroscopic methods separately and in combination with each other. The data from pollen grains and plant tissue sections can indicate variation between different samples but they also show a high heterogeneity within each sample, that can, e.g., in Raman experiments be resolved on the micron-scale. In the case of Raman microspectroscopy, substructures of pollen[13,16,31,32] and other tissue structures, such as cell walls[6,33] can be investigated. Nevertheless, for studies on the subspecies-variation, appropriate data treatment in order to take into account this heterogeneity (that leads to non-relevant variances) need to be considered. The investigation of plant cells can give insights into cell wall composition and structure and has been used for

studies on plants, particularly woods.[34–36] As another example, the investigation of biomineralization in plants has become a topic of great interest recently.[37–42] Plant cell walls consist mainly of cellulose, lignin and hemicellulose, as well as pectin.[43] The main constituents of cellulose and hemicellulose, are polysaccharides and lignin is formed by similar building blocks as sporopollenin: coumaryl, coniferyl, and sinapyl alcohols.[44] The cell wall can be divided into several layers, the primary cell wall, the secondary cell wall, and the middle lamella, that can be resolved in Raman microspectroscopy to a certain degree.[45–51] Often, research is carried out on variances within one a respective map of a plant tissue, while fewer studies analyze the variance between different maps.[7,41] In this work here, many mapping data sets are used for the comparison of specific biochemical conditions of the samples. As will be shown, for an exploratory analyis of the mapping data, a targeted extraction of subgroups of spectra is needed. As shown in previous work, spectroscopic data can be combined with other sensory data using CPCA.[52,53] The feasibility to apply such a multiblock approach to the multimodal data from the different types of plant samples will be a main aim of this work. When analyzing chemical effects that lead to small differences between data from different samples, it is very important to identify non-relevant measurement artifacts, such as the fluorescence obscuring Raman spectra, scattering in FTIR absorbance data, or superposition effects in MALDI-MS. Therefore, using different examples, advanced pre-processing methods will be discussed.

The results of this thesis are presented in Chapter 4 to Chapter 9. Chapter 2 gives an introduction about the multivariate tools and spectrometric/spectroscopic methods that were used in this thesis and summarizes recent research concerning plant spectroscopy. In Chapter 3 an overview of the samples, data acquisition, and data analysis is given. In Chapter 4, a well-designed data set of MALDI mass spectra from pollen extracts is explored. The data set is hierarchically structured comprising three different populations and four different growth conditions in each population. The variances are analyzed by principal component analysis (PCA) in combination with univariate statistical tests, and by partial least square-discriminant analysis (PLS-DA).
In Chapter 5, the variances between different populations and growth conditions within samples of the plant species *Poa alpina* will be evaluated using MALDI-TOF MS, FTIR spectroscopy, Raman spectroscopy, and SERS. The spectra will be assessed with PCA and statistical tools regarding the extent to which they are suited to solve the respective classification problems. Specifically, the combination of the data from the complementary methods and with additional plant-related information using CPCA will be presented here.
Chapter 6 addresses common benefits and challenges of Raman spectroscopy of single pollen grains. Consideration and suitable utilization of the Raman spectra of pollen that are measured with different substrates or are changed due to transport of samples are discussed. The high spatial resolution, which leads to more variance in the data sets, is particularly

noteworthy in Raman experiments. A sample set from *Sorghum bicolor* pollen, and with a complex variance, including variation between wild-type and mutant plants, but also between different growth conditions or between different breeding times is examined using Raman microspectroscopy.
FTIR-experiments on individual pollen grains are particularly challenging, because the pollen samples have to be embedded in paraffin, in order to avoid Mie scattering artifacts in the spectra.[54] In Chapter 7, different approaches are presented, how the spectra of embedded samples can be pre-processed to subsequently classify different species of pollen using chemometric methods and machine learning. Moreover, the combination of FTIR microspectra with other spectra from MALDI-TOF MS of pollen extracts and Raman spectroscopy of pollen grains using CPCA will be discussed. The reproducibility of FTIR microspectra from single pollen grains is tested using an identical sample set that was measured at a different time and by adding a data set of FTIR spectra from other pollen species.
Pollen of different plant species can be identified using MALDI imaging data of pollen extracts.[55,56] These can overlap and lead to ion suppression of individual peaks.[57] In Chapter 8, this problem is addressed by applying PLS-DA, artificial neural networks, and Random forest to identify pollen species in pollen mixtures. Particularly, using non-negative matrix factorization, the MALDI imaging data will be decomposed into components and their contribution to the mixture spectra is discussed.
In Raman mapping experiments, large maps from plant tissue can be obtained. If a data set consists of many maps for a specific experimental or physiological condition, it can be beneficial to reduce the variance before a particular step in the data analysis. Chapter 9 deals with the utilization of such large data sets in order to assess subspecies-variation, including a targeted extraction of relevant spectra. Three different examples from three different projects that focus on the silicifcation in plants are discussed here. In order to address each respective analytical question, different approaches are applied. They include the recombination of the classification results with the original spatial information, as well as the combination of the mapping data with addition information on the plants in a CPCA.
As summary of all results is provided in Chapter 10, together with an outlook on possible future directions.

2 Introduction and state-of-the-art

This chapter introduces the chemometric tools and spectrometric/spectroscopic methods that are used in this thesis. First, the chemometric approaches and algorithms are described, followed by an introduction into the methods, and subsequently, the data pre-processing will be discussed.

2.1 Multivariate statistical methods

Multivariate statistical tools are beneficial for the analytical investigation of biological systems. A biological system can be described by multiple variables at the same time. Data obtained by spectroscopic and spectrometric methods offer a large amount of variables, which can be compared to a fingerprint of the sample, since the unique spectrum is the result of a unique chemical composition. In order to evaluate the sources of variance in a data set or a sample, a large set of spectra is preferred.

Hierarchical cluster analysis (HCA)

The hierarchical cluster analysis (HCA) is an unsupervised chemometric tool that can sort a data set into clusters based on the similarities and the heterogeneity of the data. Each spectrum is a vector in a n-dimensional space, where n is the number of variables. The clustering is based on the distances between the vectors. The distance can be calculated using different metrics. The most common metric to calculate the distance is the Euclidean distance. For the two spectra p and q with n variables the Euclidean distance is:

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + ...(p_n - q_n)^2} = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \tag{2.1}$$

The heterogeneity between two spectra or sets of spectra is determined using a defined linkage criterion. Spectra with the smallest Euclidean distances to each other are clustered first and

merged afterwards with spectra or clusters with the lowest heterogeneity. The linkage that is used in this thesis is based on *Ward's* algorithm.[58] There, due to the convergence criterion, the within-cluster heterogeneity is minimized, while the heterogeneity between clusters is maximal.
The merging of clusters builds up a graph based on the heterogeneity, which is called a dendrogram. The clustering of spectra can help to understand how the samples are related to each other. In biological studies, HCA is commonly used in order to classify spectra of biological systems according to their taxonomical relation.[3,13]

Principal component analysis (PCA)

One of the most powerful tools in chemometrics is the principal component analysis (PCA). It was first mentioned in 1901 by Pearson[59] and further developed by Hotelling in 1933.[60]
PCA is used for a variance-weighted reduction of the data set. A data set can be presented as a $n \times m$ matrix, where n is the number of variables, and m is the number of spectra. Before execution, the data matrix X needs to be mean-centered, which includes the subtraction of each variable from the averaged variable. The PCA model contains four parts, the mean-centered data matrix X, the score vector t, the loading vector p and the residual matrix E:

$$X_{n\times m} = t_{1\times m} \times p^T_{n\times 1} + E_{n\times m} \tag{2.2}$$

The scores and loadings are summarized as the first principal component (PC 1). The residual matrix E can be factorized afterwards. The determined scores and loadings from the residual matrix would be PC 2. Afterwards, PC 3 can be calculated from the residuals and so on. PC 1 has the highest variance in the data set, PC 2, which is orthogonal to PC 1, represents the direction of the second highest variance. The principal components are linear combinations of the original data. Different algorithms can be applied to find the component with the maximal variance. Most commonly used are the singular value decomposition (SVD) and the nonlinear iterative partial least squares (NIPALS) algorithm by Wold[61] that is also the basic principle behind Partial-least square (PLS) (see below).
The outcome of a PCA can be represented by a scores plot. A 2D scatter plot of two different principal components visualizes how the score values from the vector t are distributed along the PCs based on the weighted variances. The spectral features causing the variance in the data set can be interpreted using the corresponding loadings p.

Consensus principal component analysis (CPCA)

Consensus principal component analysis (CPCA) is a multiblock extension of the PCA, which enables the variance weighting of variables from several data sets from different methods combined, in order to find an underlying pattern in all data. CPCA was first introduced by Wold in 1987[62] and an improved algorithm was presented by Westerhuis in 1998.[63] In CPCA, the data blocks are deflated with respect to the variation that is expressed in the so-called global scores. The detailed algorithm as described by Westerhuis[63] is presented in the following:
A CPCA with m blocks is executed using a modified NIPALS algorithm[61] starting with an arbitrary global score vector t_T. Iteratively, t_T gets optimized by first calculating the block loadings p_b and the block scores t_b, by normalization of p_b.

$$p_b = X_b^T \cdot \frac{t_T}{t_T^T \cdot t_T}; \quad \| p_b \| = 1 \tag{2.3}$$

$$t_b = X_b \cdot \frac{p_b}{\sqrt{m_{X_b}}} \tag{2.4}$$

The block score values for each block are combined to a scores matrix T.

$$T = [t_1 ... t_b] \tag{2.5}$$

Subsequently, a weighting factor w_T is estimated and normalized.

$$w_T = T^T \cdot \frac{t_T}{t_T^T \cdot t_T}; \quad \| w_T \| = 1 \tag{2.6}$$

The new global score vector t_T is estimated by

$$t^T = T \cdot w_T \tag{2.7}$$

The deflation in CPCA is conducted by equation 2.3 using the last t_T vector in order to calculate p_b, following by the subtraction from the data matrix X_b.

$$X_b = X_b - t_T \cdot p_b^T \tag{2.8}$$

A difference between performing PCA on every single block and a CPCA analysis is that the same variation appears in the same components in every data block. Furthermore, a correlation loadings plot can be generated as a result of CPCA. The data blocks that are used in CPCA are not limited to spectral data. Also, data containing discrete data as information about sensory[52] or morphological properties or even different spectral regions of the same spectra are applied in CPCA. It is not only joining the information from different data blocks in one analysis, it also enables the evaluation of interactions between the different blocks.[64,65]

Partial least square discriminant analysis (PLS-DA)

Partial least square discriminant analysis (PLS-DA) is the classification approach of the commonly used multivariate regression PLS regression (PLS-R). The algorithm was introduced by Wold based on the NIPALS algorithm.[63,66,67] PLS-R is a regression method using a vector, PLS-DA is applied on a target matrix Y. The main difference between the multivariate regression method PLS and PCA/CPCA is the application of the matrix factorization not only on the data matrix X, but in addition also on Y. Therefore, the factorization can be defined as

$$X_{n\times m} = t_{1\times m} \times p^T_{n\times 1} + E_{n\times m} \tag{2.9}$$

For the decomposition of the X matrix and

$$Y_{n\times m} = u_{1\times m} \times q^T_{n\times 1} + F_{n\times m} \tag{2.10}$$

for the Y matrix, respectively.
The iterative algorithm can be described as follows:[63] An arbitrary scores vector u is chosen as a starting point and is finished when convergence of the scores vector t is reached. First, the weight w of the data matrix X is calculated in order to estimate the scores t in X.

$$w = X^T \cdot \frac{u}{u^T \cdot u}; \quad \| w \| = 1 \tag{2.11}$$

$$t = X \cdot \frac{w}{w^T \cdot w} \tag{2.12}$$

Using the score vector t the weight q and score vector u of the Y-vector/matrix can be obtained.

$$q = Y^T \cdot \frac{t}{t^T \cdot t} \tag{2.13}$$

$$u = Y \cdot q \tag{2.14}$$

The deflation is induced by the estimation of the loading vector p on X using the last score vector t.

$$p = X^T \cdot \frac{t}{t^T \cdot t} \tag{2.15}$$

Subsequently, the data matrix X and the regression matrix Y are getting decomposed.

$$X = X - t \cdot p^T \tag{2.16}$$

$$Y = Y - t \cdot q^T \tag{2.17}$$

PLS is one of the so called supervised methods since a training set is needed as an input to build a regression model. The model needs to get evaluated using suitable validations methods, e.g., an external validation. To achieve this, the model is trained and tested with two different data sets. In addition, regression parameters b and b_0 can be calculated. The regression equation is:

$$y = b_0 + Xb \tag{2.18}$$

X can be replaced using the equations described above. The regression parameter are calculated with:

$$b = W(P^T W))^{-1} q \tag{2.19}$$

and

$$b_0 = \bar{y} - \bar{x}^T b \tag{2.20}$$

Non-negative matrix factorization

Non-negative matrix factorization (NMF) decomposes the data matrix X into k components and their contribution to the data matrix. As an outcome, the components and contributions are positive. NMF was introduced by Lee and Seung in 1999 and was applied for text mining and pattern recognition.[68] Nowadays, NMF is widely used for unsupervised data reduction and as an exploratory tool in data science.[69] Specifically, NMF can help to understand underlying patterns in spectral data.[45,70–72] Components and their contributions are calculated using Low rank approximation, e.g., singular value decomposition.[73] For NMF this approach requires the number of ranks k, that can be chosen by trial and error or using *a priori* information of the data.

Artificial neural networks

Artificial neural networks (ANN) are commonly used as a machine learning approach for pattern recognition. In principle, ANN follows the idea of the neural network of the human brain,[74] where, e.g., for spectral data, the input is referring to the variables of the data and the output is defined by the affiliated classes. The architecture of the network, i.e. how input and output are connected, influences the outcome and can be optimized for the specific classification problem.[75] Here, feed-forward neural networks are applied. The network contains three layers, the input layer, the hidden layer, and the output layer. The input neurons, e.g., the Raman-shifts or wavenumbers, are weighted and combined while training the model and merged into the hidden neurons and subsequently into the output neurons. ANN is beneficial since they enable the classification of non-linear variance within the data.

Random forest

Random forest (RF) is a classification model based on many decision trees. It was first proposed by Breiman in 2001.[76] A decision tree in general consists of several branches, where a classifier, e.g., a threshold, assigns data regarding a specific class.[77] In Random forest, the results of the decision trees are combined together, which is called bagging. Each decision tree is generated with a random subset of the data's variables and is creating the random forest, where each decision tree contributes to the overall classification.[78]

2.2 Spectrometric and spectroscopic methods as analytical tools to investigate biological systems

Spectrometric and spectroscopic methods have been used for decades to study plant material.[33,79,80] They are often combined with chemometric tools to classify and characterize the spectra or to explain the source of variances within a set of spectra with respect to a certain biological affiliation, e.g., into different pollen species[2,12,13] or cell wall substructures.[47,81,82] The following section presents the four spectrometric and spectroscopic methods which are the most relevant for the present thesis, matrix-assisted laser desorption/ionization mass spectrometry (MALDI), Fourier-transform infrared (FTIR) spectroscopy, Raman spectroscopy, and surface-enhanced Raman scattering (SERS).

Matrix-assisted laser desorption/ionisation mass spectrometry

Matrix-assisted laser desorption/ionization (MALDI) is based on the co-crystallization between the molecules of a sample and a matrix. The sample is mixed with a matrix compound, such as α-cyano-4-hydroxycinnamic acid (HCCA), and poured on a target. The dried co-crystals get irradiated with a laser, which leads to energy absorption by the matrix and to evaporation and ionization of the analyte. Subsequently, the molecules are detected by a mass spectrometer (MS), e.g., a time-of-flight analyzer (TOF) and mass spectra with the intensity as a function of the mass-to-charge ratio can be obtained.
MALDI enables soft ionization of molecules, which means that large biomolecules remain intact or show low fragmentation. As a result, the mass of intact biomolecules such as lipids[83,84] and proteins[85,86] can be studied. Therefore, MALDI mass spectra provide mass spectrometric fingerprints, that can be used to characterize and classify complex biological samples such as bacteria,[87–91] fungi,[84,92–97] and plant tissues.[98–102]
Specifically, MALDI-TOF MS is suitable to investigate allergenic molecules in pollen.[103–106] In combination with gel electrophoresis, the specific allergens can be extracted from the pollen and identified by comparing the masses of the proteins with those of online database entries.[104–106] Moreover, MALDI mass spectrometry can help to classify pollen from different plant species.[1,2,107] Particularly, studies shown that MALDI-TOF MS enables robust classification of pollen spectra in combination with chemometric methods, such as hierarchical cluster analysis (HCA)[1] and principal component analysis (PCA).[2,107] The identification of different pollen species was achieved for pollen grains trapped on sticky carbon tape as a simulation of a common pollen trap using MALDI-TOF MS and chemometrics.[107]

In mass spectrometry imaging (MSI), the high specificity of MALDI MS is used to visualize the distribution of certain molecules, e.g. bio markers in tissues. Usually, the image is

created by mapping the intensity of one peak or the ratio of two peaks that are specific for a compound in the sample.[108] In recent years, several studies using MSI on plant tissues have been published[109–111] and it was also used for the identification of pollen species in mixtures.[56]
Mass spectrometric images can also be analyzed using chemometric approaches, such as HCA and PCA,[71,112] or by matrix factorization methods such as NMF.[71]

FTIR spectroscopy

IR spectroscopy is based on the interaction between molecules and light in the frequency range. Molecules absorb the IR light which results in their excitation to higher vibrational energy levels. The wavenumber range can be divided into the far IR (FIR, 400-50 cm^{-1}), the mid IR (MIR, 4000-400 cm^{-1}) and the near IR (NIR, 14000-4000 cm^{-1}). For the analysis of biological systems, the most relevant region is the MIR, since the vibrational bands of biomolecules are observed in this region of the IR spectrum. The region between 500-and 1500 cm^{-1} is also called the fingerprint region.
FTIR spectroscopy has become an established tool in plant biology[44,113–117] and in particular, it also shows potential in the discrimination of different pollen species.[4,5,9,10,118–122] In addition, FTIR spectroscopy allows the classification and characterization of chemical variation at the sub-species level in pollen, specifically between populations of the same species, and lead to conclusions regarding the phenotypic plasticity within plant populations, i.e., how the plant can adapt to external influences.[14,15,30,123]
The majority of FTIR pollen studies were conducted by measuring bulk pollen samples[4,14,15,124,125] and only few studies exist about single pollen grain measurements,[10,32,54,118,126] due to scattering artifacts that can occur in MIR spectra of micron-scaled size particles.[127] The spectral contribution from Mie scattering can superimpose the absorbance spectrum, depending on the geometry of the sample, and it can cause band shifts, distortions and artificial bands.[127–130]
These scattering problems can be addressed by modification of the experiment, such as multi-grain measurements with large aperture[10,118,126] or measurement in an embedding matrix.[54]
Specifically, appropriate data pre-processing can reduce the influence of the scattering artifacts in the spectra, such as extended multiplicative scattering correction (EMSC)[131] (For more details see below in this chapter).

Raman spectroscopy and Raman mapping

When a molecule interacts with monochromatic light, scattering occurs that can be either elastic (Rayleigh scattering) or inelastic (Raman scattering). The effect of the inelastic scattering

was discovered independently by Raman and Krishnan[132] and by Landsberg and Mandelstam[133] in 1928 and first predicted by Smekal in 1923.[134] Caused by the interaction of the molecule with incident light, the molecule is excited to a virtual energy level. Depending on whether the molecule is in the ground state or the vibrational excited state, Stokes or Anti-Stokes scattering can occur, where the Stokes bands have a higher intensity.
The selection rules of Raman spectroscopy dictate that a vibration is Raman active if the polarizability of the molecule changes with the vibrational mode. Therefore, Raman spectroscopy provides complementary information to FTIR, where the selection rules are based on changes in the dipole moment of the vibrating molecule. Because of the selection rules, IR spectroscopy suffers from the high absorption of water, while Raman spectroscopy has the advantage of lower water contribution in the spectra. Therefore Raman spectroscopy is a promising method for the study of biological samples, such as cells,[135–137] microorganism,[138–140] histological investigation of animal tissues and plant material,[47,141–144] where in particular the investigation of pollen samples should be pointed out.[11–13,16,145,146] The investigation of pollen grains using spectroscopic methods has become increasingly important in recent years.[10,16,32,121,124,147–150] Raman spectroscopy can be applied as a quick identification method for pollen warnings.[13] Recently, a high throughput Raman spectroscopy approach was applied for the successful discrimination of pollen.[148] In addition, the chemical fingerprint in the Raman spectrum can provide information about the chemical composition of pollen grains and the changes in the composition caused by different substructures of the pollen grain or even by external factors.[16,31]
Raman spectroscopy is often combined with chemometric methods in order to classify and characterize the Raman spectra according to a biological question. Most prominent examples are the implementation of hierarchical cluster analysis (HCA)[5,12,13] and principal component analysis (PCA).[5,16,124]

Due to the non-invasive and non-destructive nature of Raman microscopic experiments and the micro-scale spatial resolution, Raman imaging became a valuable approach for the study of plant tissues.[7,32,45,46,48,50,81,151–153] Similar to MSI (see above) an image is created by scanning the tissue and assigning one spectrum to a specific xy-position each. Therefore, the chemical composition of different compartments, as well as the substructure of the tissue can be studied.[6,35,36,151]
The Raman images can be analyzed using univariate approaches, by taking, e.g., the intensity of a certain band of each xy-position in a sampled area. In addition, multivariate approaches can help to elucidate the chemical composition of tissues.[7,32,46]

Surface-enhanced Raman scattering

The cross sections of Raman scattering can be low, which leads to low signal-to-noise-ratio in the spectra. Depending on the sample, higher exposure time would be necessary. In addition, fluorescence contributions occur in the spectra of biological samples and can mask bands. This phenomenon has been described for Raman spectra of pollen in the literature.[11,12,146] Several solutions have been developed to reduce the fluorescence contribution by either using another excitation wavelength or photobleaching the sample.[13]
In another approach, fluorescence can be quenched by adding metal nanostructures to the sample, which results in the enhancement of the Raman signal intensity.[19,154] The enhancement of the Raman signals using metal surfaces is known as surface-enhanced Raman scattering (SERS). The phenomenon was first observed by Fleischmann *et al.* in 1974, when they received high Raman scattering of pyridine on a silver electrode surface.[155] The enhancement in SERS can be several orders of magnitude.[156] This can be attributed to two different mechanisms: chemical and electromagnetic enhancement.
Similar to a normal Raman experiment, the nanostructures are irradiated with light of a certain energy. This leads to excitation of the localized surface plasmon modes of the nanoparticle (collective oscillations of the electron cloud), when they are in resonances with the incident wavelength. If the scattering emitted by analyte molecules is also in resonance with the plasmons, the field is also amplified and the signals get enhanced. The intensity of the SERS signals increases with decreasing distance between the molecules and the nanostructures.
In addition, the electronic coupling between the metal surface and the molecule leads to a broadening of the electronic states,which results in an increase in the polarizability of the molecule. This process is also known as chemical enhancement.
SERS is frequently used to study biological systems. Specifically, living cells are studied using SERS.[157–161] Unfortunately, the enhancement decreases drastically with increasing distance between molecule and nanoparticle. Especially, the investigation of complex analytes in solution is challenging. Seifert *et al.* showed a SERS application, where a high amount of spectra is needed in combination with an appropriate selection of relevant spectra in order to analyze the data.[3] SERS is often combined with chemometric tools in order to classify and characterize biological systems.[162–165] In particular, pollen can be investigated using SERS.[3,166] Joseph *et al.* studied the pollen outer shell using silver nanoparticles.[19] The taxonomic relationships of the pollen can be examined using the water-soluble components of the pollen grains.[3] Due to the complementarity to Raman spectroscopy of intact single pollen grains, SERS of the water-soluble parts in pollen enables additional spectroscopic information for multiblock analysis.

2.3 Pre-processing routines for spectral analysis

Before applying chemometric methods to the data sets, variances that are caused by technical artifacts of the respective spectroscopic/spectrometric methods should be minimized. In addition, spectra can suffer from influences, e.g., fluorescence in Raman spectra, or Mie scattering in infrared absorbance data that may not add directly to the understanding of the sample. Such effects also need to be addressed before data analysis. The pre-processing steps have to be specific for the kind of data and the biological questions of the experiment. In general, most of the data used in this thesis can be pre-processed by the three following steps, interpolation, baseline correction and normalization of the data. Interpolation is applied on mass spectra and Raman spectra, including SERS spectra.
Regarding the mass spectra the amount of variables of each spectrum can be high (e.g. ~30000 for the *m/z* 1000 - 12000). The computational cost of analyses is increasing exponentially with the data size, which can lead to longer calculation times.[167] A common approach is to reduce the spectrum to a list of peaks (*m/z* and intensity-pairs) for the analysis, but this would result in data with different numbers of variables, which would not be suitable for the chemometric approaches proposed for the analyses discussed in this thesis.
In the case of Raman spectra, obtained with dispersive instruments, interpolation ensures a uniform distance between two data points across the entire spectrum and thus prevents an increased weighting of variables, as typically, the number of data points varies across the spectral range. The optimal distances between data points can be estimated by averaging the distances of the data points within different spectral ranges (e.g., 400-600 cm^{-1} and 1600-1800 cm^{-1}). For example, if the distances are ranging from 1.3-1.6 cm^{-1}, an optimal distance of 1.45 cm^{-1} can be applied.
Using interpolation, a function is estimated using the spectra and a defined set of junction points. As a result, the data are interpolated with equal distances and in a defined spectral range.[168] The estimated function is based on the junction points that define the original data points.

In all spectral data sets in this work, one has to deal with unwanted distortions in the spectra, i.e. the 'background'. For each method a different source of the background is known:
I) In a MALDI-TOF MS experiment, the background is mostly caused by the desorption and ionization of the applied matrix and impurities.[169]
II) In FTIR spectroscopy, the background can be a result of scattering effects, additional absorption by the substrate, e.g., the ZnSe-slide, or by the instrumental conditions.[170]
III) Many molecules cause fluorescence in Raman spectra resulting in a strong background that can mask bands, particularly in Raman spectroscopy of biological samples.[171]
Researchers use different approaches for digital baseline correction such as a polynomial fit of a higher degree,[172,173] Savitzky-Golay smoothing[174] or wavelet correction.[175] For Raman

and FTIR spectra, the first and second derivatives are often calculated before data analyses, in order to remove offset and slope of the baseline.[176]
Here, the spectra were baseline corrected using asymmetric least square (AsLS) smoothing, as proposed by Eilers.[177,178] The AsLS algorithm is an extension of the Whittaker smoothing suggested in 1922.[179] It has been shown, that the algorithm works well on different types of data[178,180–185] and, therefore, AsLS background correction is applied to all kinds of spectral data shown in this thesis.

Intensity normalization of the data is necessary since variations based on physical contributions among spectra need to be minimized before data analysis, such as classification. Especially during measurements of biological samples, individual differences in sample thickness as well as in other preparation steps lead to separation of spectra. In addition, instrumental conditions can fluctuate during measurements. Due to the Mie scattering in FTIR spectroscopy of micron-sized samples a normalization regarding the absorbance as well as the band positions need to be carried out.
Usually, a spectrum is normalized by multiplying a scale factor to each point of the spectrum. The scale factor can be for example the sum of each point in the spectrum. This normalization, which is commonly applied for mass spectra is known as total ion count (TIC). Vector normalization, also called 2-norm is based on the Euclidean norm of the data points (v_i) of the spectra. The 2-norm of a spectrum is defined by equation 2.21.

$$v = \sqrt{\sum_{i=1}^{n} (v_i)^2} \tag{2.21}$$

For discrete data, e.g., a list of morphological properties of the sample, autoscaling was applied as normalization. Therefore, a data set X with n spectra and i variables were mean-centered and subsequently divided by the standard deviation σ.

$$X_{mean} = X - \frac{1}{n}\sum_{i=1}^{n} X_i \tag{2.22}$$

$$X_{autoscaled} = \frac{X_{mean}}{\sigma_{X_i}} \tag{2.23}$$

Such a normalization is dividing the spectra by a scaling factor. Especially, in MALDI investigations of biological samples, the suppression effects[57] need to be considered and a more complex model based normalization is suggested.[186]
Extended multiplicative scattering correction (EMSC) is a model-based normalization commonly applied for FTIR spectra.[130,187] The main application has been the minimization of

scattering effects in the spectra that occur in measurements of micron-sized particles.The basic principle of EMSC is the separation of the spectral, that is, chemical information and spectral distortions that are often the result of physical effects, such as scattering. The model of the an absorbance $A(\upsilon)$ can be described by 2.24.[188]

$$A(\upsilon) = a + z_{ref}(\upsilon) \cdot b + d_1\upsilon + d_2\upsilon^2 + ... + d_n\upsilon^n + e(\upsilon) \tag{2.24}$$

where z_{ref} is either a reference spectrum or the average spectrum of the data set[131,188] and a,b to $d_n\upsilon$ describe the modeled parameters, and $e(\upsilon)$ the residuals. The basic EMSC includes linear and quadratic terms. The corrected spectrum would be estimated by 2.25.[188]

$$A(\upsilon)_{corrected} = \frac{A(\upsilon) - a - d_1\upsilon - d_2\upsilon^2}{b} \tag{2.25}$$

3 Materials and Methods

This chapter introduces the materials and methods used in this thesis. First, the samples and corresponding data sets are listed. Subsequently, the data acquisition is presented, followed by the descriptions of the multivariate data analyses.

3.1 Plant material and sources of data

Different sample/data sets from plant tissues were used for the studies. The data were obtained by own experiments and in collaboration by co-workers as part of different projects.

3.1.1 Data from pollen samples

Most of the analyses were executed on data of pollen grains to investigate differences between and within pollen species. Different sets of pollen samples were assembled to explore different biological questions and assess the subspecies variances within hierarchical (Chapter 4 and Chapter 5) and complex (Chapter 6) data sets. Some of the samples are reused in different sample sets. The sample/data sets were named after their sampling location.

Pollen Norway I

This sample/data set is discussed in Chapter 4 and Chapter 5. Measurements of representative samples from the set were presented in Section 6.1. The sample set Pollen Norway I contains 272 samples (one sample for each plant) of the three grass species *Poa alpina*, *Anthoxanthum odoratum*, and *Festuca ovina*. For an assessment of the differences in the chemical composition of pollen species, plants of three populations were studied in *Anthoxanthum odoratum* and *Poa alpina*. In addition, one population of *Festuca ovina* was investigated. The greenhouse experiments were designed and executed by Prof. Dr. Siri Fjelheim, the sampling

was conducted by Dr. Boris Zimmermann and Dr. Murat Bağcıuğlu from Norwegian University of Life Science (NMBU). This sample set is part of a larger set analyzed and discussed by Zimmermann *et al.*[15]
Seeds of the populations were chosen to cover climatic and geographic variances and were acquired from the Nordic Gene Bank. The populations of *Anthoxanthum odoratum* come from France, Greece, and Finland. Those of *Poa alpina* originate from Sweden, Italy, and Norway. The *Festuca ovina* population was from Sweden. The individuals grew outside over summer, after which each individual was divided into four clones. The plants were subsequently vernalized for 12 weeks at 4 °C with a day length of 8 hours. After vernalization, both the temperature and the nutrient addition were varied, so that the clones grew under four different conditions. Day length was increased to 16 hours to induce flowering. The plants were grown at high temperature (20 °C) or low temperature (14 °C) in combination with additional nutrients in the irrigation water (+ nu) or no additional nutrients in the irrigation water (- nu).[15] The detailed scheme of the sample set is given in Chapter 4 (Figure 4.1). The pollen set Norway I was measured using matrix-assisted laser desorption/ionization time of flight mass spectrometry (MALD-TOF MS).

For comprehensive multimodal study, a subset of Pollen Norway I containing 72 samples of the three populations from *Poa alpina* was measured also using FTIR spectroscopy, Raman spectroscopy, Surface enhanced Raman scattering (SERS), and MALD-TOF MS. An overview of the samples can be found in Figure 5.1 in Chapter 5. All data were obtained in own prior projects (Raman spectroscopy, SERS, and MALD-TOF MS[189]) and in addition by co-workers Prof. Dr. Achim Kohler, Dr. Boris Zimmermann, and Dr. Murat Bağcıuğlu from NMBU (FTIR spectroscopy, additional plant data).[15]

Pollen Israel

The sample set Pollen Israel consists of 25 pollen samples from 25 individual *Sorghum bicolor plants*. In Section 6.2. the variance of this data set is evaluated. The plants were growing in a greenhouse in Rehovot, Israel by collaboration partners Fikadu Biru, Prof. Dr. Rivka Elbaum and Nerya Zexer from the Hebrew University Jerusalem.
Sorghum bicolor BTx623 wild-type and two different types of mutants, namely *sblsi1*,[37] where the silicon transporter Lsi1 is deactivated and *bmr* (brown midrib mutants) with low lignin content in cell walls,[190] were grown in a greenhouse (The Hebrew University, Faculty of Agriculture, Rehovot, Israel) for two months (sblsi1 mutants) and three months (*bmr* mutants) under natural light and temperatures.
Particularly, growth experiments regarding different stress conditions were carried out by Fikadu Biru. Four wild-type plants and four *sblsi1* mutant plants were treated under drought

or under salt stress conditions, respectively.[191] In summary, drought stress was induced by exposing the plants to a shortage of water for two weeks. Salt stress was initiated by adding different concentrations of sodium chloride to the irrigation water. A detailed description of the growth experiment is published by Biru.[191]
Pollen from the flowering plants was collected in collaboration with Fikadu Biru, Nerya Zexer, and Victor Rodriguez-Zancajo. A schematic presentation of the data set is shown in Chapter 6 in Figure 6.9 The sample set is measured using Raman microspectroscoy and MALD-TOF MS.

Pollen Norway II

The sample set Pollen Noway II comprises in total 71 pollen samples from nine different Poaceae species. The pollen samples were obtained by the Norwegian co workers Prof. Dr. Siri Fjelheim, Dr. Boris Zimmermann, and Dr. Murat Bağcıuğlu, who did the growth experiments and the pollen sampling respectively. The set Pollen Norway II contains pollen samples of two populations each from *Poa alpina* (Sweden and Italy) and *Anthoxanthum odoratum* (France and Greece) that were already described above.
For different purposes, the data set is split into Pollen Norway IIa and Pollen Norway IIb regarding the different experiments. The samples set Pollen Norway IIa is discussed in Section 7.1-7.4, the sample set Pollen Norway IIb in Section 7.5, respectively. A schematic presentation is given in Figure 7.1. Pollen samples were collected from two populations from each of the five Poaceae species *Anthoxanthum odoratum*, *Bromus inermis*, *Hordeum bulbosum*, *Lolium perenne*, and *Poa alpina* and from one population of the four pollen species *Hordeum vulgare*, *Hystrix patula*, *Piptatherum millaceum*, and *Piptochaetium avenaceum*. From each population, up to five individuals of different genotypes were used in the experiment (six samples for *Hystrix patula*).
Bromus inermis, *Hordeum bulbosum*, *Hordeum vulgare*, *Hystrix patula*, *Lolium perenne*, *Piptatherum millaceum*, and *Piptochaetium avenaceum* plants were grown in an open greenhouse at 17 °C for four weeks. Subsequently, temperature (vernalization at 4 °C for six weeks and then transferred to 17 °C , or no vernalization at 17 °C) and day lengths (8 or 16 hour photoperiod) were varied for different plants as required by another study from which the plants were sampled.
Pollen were collected from the plants at the onset of pollination (varying for each species and growth condition) and stored at -20 °C . The sample set Pollen Norway IIa is measured by FTIR microspectroscopy and MALD-TOF MS. In addition, data were obtained by Raman microspectropy experiments conducted by Simon Schröder as part of a research internship (unpublished data).

Pollen Berlin

The data set Pollen Berlin contains MALDI mass spectra of pollen from 16 different plant species, that were obtained by Dr. Fransiska Lauer at BAM, Berlin Germany. The data set is discussed in Chapter 8. Pollen samples of *Artemisia absinthium*, *Betula occidentalis*, and *Populus nigra* were purchased from Sigma Aldrich (Sigma-Aldrich, St. Louis, MO, USA). Pollen samples from *Alnus cordata*, *Alnus rubra*, *Betula alleghaniensis*, *Betula ermanii*, *Betula tatewakiana*, *Corylus avellana*, *Corylus sieboldiana*, *Philadelphus californicus* , *Pinus mugo*, *Philadelphus pubescens*, *Pinus rigida*, *Pinus sylvestris*, and *Syringa reticulata* were collected in the botanical garden in Berlin by Dr. Franziska Lauer in 2016. All pollen samples were stored at -20 °C until measurement.
Two MALDI imaging data sets (mixture 1 and mixture 2) are analyzed in Chapter 8 of this thesis. The first imaging data set consists of 154 spectra of an artificial pollen mixture of *Artemisia absinthium*, *Betula occidentalis*, and *Populus nigra*, the second imaging data set has 136 mixture spectra of *Alnus cordata*, *Corylus avellana*, and *Pinus sylvestris*.
In addition, a database of reference mass spectra were used, including varying amount of mass spectra for each species. The content of the database is shown in Table 3.1.

Table 3.1: Amount of spectra from the database in the data sets of Pollen Berlin.

Pollen species	amount of spectra	Pollen species	amount of spectra
Mixture 1	154	*Corylus avellana*	294
Mixture 2	136	*Corylus sieboldiana*	212
Artemisia absinthium	32	*Philadelphus californicus*	240
Alnus cordata	231	*Pinus mugo*	102
Alnus rubra	61	*Populus nigra*	32
Betula alleghaniensis	243	*Philadelphus pubescens*	193
Betula ermanii	181	*Pinus rigida*	187
Betula occidentalis	30	*Pinus sylvestris*	85
Betula tatewakiana	103	*Syringa reticulata*	177

3.1.2 Data of plant tissue sections

Data sets of plant tissues used in this study are obtained by co-workers Prof. Dr. Rivka Elbaum, Dr. Zsuzsanna Heiner, Dr. Sabine Holz, Ingrid Liedtke, and Nerya Zexer. All data are obtained by Raman imaging experiments and discussed in chapter 9.

Cucumber

The Cucumber data set contains Raman imaging data from *Cucumis sativus* Sonja. The plants were grown in a greenhouse in Berlin, Germany by collaboration partner Dr. Sabine Holz. Details of the growing experiments are published by Zeise *et al.*[7] In summary, seeds of *Cucumis sativus* Sonja were cultivated in silicic acid poor soil. Two cucumber plants were irrigated with tap water containing silicic acid (2 mM) and adjusted to pH 7 and two plants were irrigated with tap water without additional silicic acid.[7,192]
The cross sections were taken from different plant organs. For each of the four plants four to seven cross sections were cut by Ingrid Liedtke for the plant organs respectively and stored at 4 °C experiments. More details about the sample preparation are described by Zeise *et al.*[7] A schematic presentation of the data set is shown in Chapter 9 in the Figure 9.7.

Sorghum

Raman imaging data from Sorghum plant *Sorghum bicolor* (L.) Moench tissues were studied. Growing experiments and preparation of the cross sections were executed by Ingrid Liedtke. Four plants of *Sorghum bicolor* (L.) Moench (line BTX-623) and four *SbLsil*- mutant[37,193] at room temperature as described in more detail in reference [152].[152]
Cross sections were collected from the 7th leaf of each plant following the numbering of Kumar *et al.*[193] A schematic presentation of the data set is shown in the respective chapter in the Figure 9.12.

One Raman map of Sorghum root tissue was obtained by Nerya Zexer.
Seedlings of *Sorghum bicolor* (L.) Moench (line BTx623) were grown hydroponically for one week. Cultivation was carried out in a growth chamber under controlled conditions. A detailed description of the growing experiment is published by Zexer *et al.*[40]

3.2 Data acquisition

3.2.1 Sample preparation and data acquisitions of MALD-TOF MS measurements

The data set Pollen Norway I, as well as 70 samples from Pollen Norway II and 14 samples from the data set Pollen Israel were obtained by MALD-TOF MS following an established protocol.[1,2] The pollen samples were deposited directly on a MALDI stainless steel target

(MTP 384). Measurements of the data set Pollen Berlin were conducted by Dr. Franziska Lauer.
The reference spectra of the 16 different Pollen species were acquired performing MALD-TOF MS after a protocol by Lauer *et al.*[107] The pollen grains were deposited on sticky carbon tape (P77817, Science Services GmbH, Munich, Germany) on a MTP 384 standard target.
On each sample, 1 μL of formic acid (90 %) was pipetted for extraction. After drying at room temperature, 1 μL of matrix solution (10 mg of α-cyano-4-hydroxycinnamic acid in 1 mL 1:1 acetonitrile/water and 0.1 % trifluoroacetic acid) was added to each spot, and the target was left to dry at room temperature.
The two imaging data sets of two different pollen mixtures were obtained from pollen fixed on sticky carbon tape before adding the matrix (α-cyano-4-hydroxycinnamic acid in 1 mL 1:1 acetonitrile/water and 0.1 % trifluoroacetic acid). An Autoflex III MALD-TOF mass spectrometer (Bruker Daltonik, Bremen, Germany) in positive linear mode and with a 355 nm Smartbeam laser (200Hz) was used for all measurements. Spectra in the mass range from m/z 1000 to 15000 were recorded for each sample. In order to account for potential heterogeneity within the sample spots, 2000 spectra from four different positions of each spot were accumulated.
For acquisition of the two MALDI imaging data sets, the software FlexImaging (Bruker Daltonik, Bremen, Germany) was used. The MALDI target was scanned using a pixel size of 100 $\mu m \times 100$ μm and 1000 spectra were accumulated for each spot.[55]

The spectra were interpolated using the in-built *interp1*-function in Matlab with a distance of m/z 2, in order to reduce the data set size. Afterwards, the spectra were baseline-corrected using asymmetric least square correction as proposed by Eilers[177,178] and vector normalized before further data analysis.

3.2.2 FTIR spectroscopy

The samples of Pollen Norway I were measured using FTIR spectroscopy by the the co-workers Dr. Boris Zimmermann and Dr. Murat Bağcıuğlu. The samples were prepared as bulk samples, following the protocol published by Bağcıuğlu.[123] 1 mg of each pollen sample was centrifuged with 500 μL of distilled water using a 2 mm probe coupled to a Q55 Sonicator ultrasonic processor (QSonica, LLC, USA) under 100 % power. Afterwards, the pollen suspension was centrifuged with 13,000 rpm for 10 min, and 400 μL of supernatant was removed from each sample. Three aliquots of 8 μl each were pipetted onto an IR-transparent silicon 384-well microtiter plate (Bruker Optik GmbH, Germany) and dried at room temperature for 1 h.
FTIR measurements were executed using a HTS - XT extension unit coupled to a TENSOR 27 spectrometer (both Bruker Optik GmbH, Germany). The system is equipped with a globar

mid - IR source and a DTGS detector. The spectra were recorded in transmission mode, with a spectral resolution of 4 cm^{-1} and digital spacing of 0.964 cm^{-1}. A spectrum from an empty well of the microtiter plate was recorded as background spectrum before each measurement. Spectra were obtained using an aperture of 5 μm and 32 scans for each spectrum.
Spectra were pre-processed using extended multiplicative signal correction (EMSC) with linear and quadratic components[131,174,194] on the second derivatives with a polynomial of degree two and a window size of 7 points.[174] The spectral range from 800-1800 cm^{-1} was used for multivariate analysis. An average spectrum was calculated from the spectra of the three technical replicates (aliquots) per sample, resulting in a set of 72 average spectra for multiblock analysis. The results are discussed in Chapter 5.

The data set Pollen Norway II (71 samples corresponding to 71 plants) was measured using FTIR microspectroscopy. Therefore, a procedure involving paraffin embedding proposed by Zimmermann *et. al*[14] was adapted. The analysis of the data set is discussed in Chapter 7 of this thesis.
For FTIR microspectroscopy, the pollen grains were spread onto a thin layer of paraffin on a ZnSe slide. With the help of a glass slide, the soft paraffin (Enzborn Vaseline, Nordwalde Germany) was distributed over the pollen grains, resulting in embedding of the pollen grains in the thin paraffin layer.
FTIR spectra were obtained in transmission mode using a Nicolet FTIR microscope (Thermo Scientific, Waltham, USA), equipped with a single element MCT detector and with a 32x Cassegrainian objective. The size of the sampled spot was 15 μm x 15 μm. As light source, a synchrotron source (beam line IRIS, HZB-BESSY, Berlin) was used. The FTIR spectra were measured with a spectral resolution of 4 cm^{-1} and digital spacing of 1.9 cm^{-1}, by averaging 128 interferograms per spectrum. A background spectrum was collected from the ZnSe slide with identical parameters. For each of the 71 samples, approximately 20 different pollen grains were measured per plant (with one spectrum per pollen grain), resulting in a data set of 1004 spectra in total.
For each plant sample, 2 to 5 spectra of the pollen-free paraffin layer were measured using the same condition as described above, leading to 190 pure paraffin spectra. Finally, individual pollen grains were measured on a ZnSe slide without paraffin embedding (i.e. unembedded samples). Approximately 20 spectra of individual pollen grains from only one plant per grass species were measured, resulting in 97 spectra of unembedded pollen grains in total.
The spectral pre-processing steps were adapted regarding the application of four different approaches in order to utilize the pollen spectra with paraffin contribution. The four approaches (approach 1, without further consideration, approach 2, omitting affected spectral range, approach 3, extraction of paraffin contribution using non-negative matrix factorization, approach 4, diminishing paraffin variation using a complex EMSC model) are discussed in Chapter 7. In general, the pre-processing steps include baseline correction as proposed by

Eilers,[177,178] normalization, either vector normalization or EMSC[131] and the calculation of average spectra with respect to the following data analysis. All pre-processing steps with respect to the approach is summarized in Figure 7.5 in Chapter 7. All spectral pre-processing was performed using Matlab (MathWorks, Inc., 2015b).

3.2.3 Raman microspectroscopy

The sample sets Pollen Norway I, Pollen Norway IIa (in collaboration with Simon Schröder), and Pollen Israel were measured by Raman spectroscopy. In addition, plant cross section of the data sets Cucumber and Sorghum cross sections were obtained by Raman imaging data by collaboration partners Dr. Zsuzsanna Heiner, Ingrid Liedtke, and Nerya Zexer.
Raman spectra for the data sets Pollen Norway I (discussed in Chapter 5), Pollen Norway IIa (by Simon Schröder) (discussed in Section (7.4.), and Pollen Israel (discussed in Chapter 6.2) were obtained from single pollen with a LabRam HR microspectroscopic setup (Horiba Jobin-Yvon GmbH, Bensheim, Germany) equipped with a liquid nitrogen-cooled CCD detector and a 50x long distance objective (Olympus, Hamburg Germany) with a numerical aperture of 0.55.
72 pollen samples of *Poa alpina* from the data set Pollen Norway I were measured during the period of the master thesis[189] and reused for a more comprehensive multiblock analysis, here (Chapter 5). The pollen grains were measured using a diode laser operating at a wavelength of 785 nm and an intensity of $7 \cdot 10^5 W/cm^2$. For each sample, ten spectra from ten different single pollen grains were collected, using an accumulation time of 10 s per spectrum. In total, 720 individual spectra were obtained, resulting in 72 average spectra for multiblock analysis.

Furthermore, Raman experiments were conducted on three samples of Pollen Norway I, one for each grass species of *Poa alpina*, *Anthoxanthum odoratum*, and *Festuca ovina* (Section 6.1). The samples were measured on calcium fluoride and, in addition, two samples from *Anthoxanthum odoratum* and *Festuca ovina* were measured on carbon tape, using a Nd:YAG diode-pumped solid state laser with a wavelength of 532 nm and a filter of 1 % (intensity 3.1 $*10^4\ Wcm^{-2}$ on the spot) and spatial distance of 5 μm.
The 25 samples (approx. 5 individual pollen grains) from the sample set Pollen Israel (discussed in Section 6.2) as well the 50 samples from Pollen Norway IIb (Simon Schröder (discussed in Section 7.4)), 10 individual pollen grains for each sample/plant) were obtained using a laser with a wavelength of 785 nm (intensity $1.4 \cdot 10^6\ Wcm^{-2}$) and spatial distancing of 2 μm. All spectra were obtained with an accumulation time of 1 s and recorded over the spectral range from 350 cm^{-1} to 1750 cm^{-1}.
For frequency calibration, six bands in the spectrum of 4-acetaminophenol (1648.4, 1323.9,

1168.5, 857.9, 651.6, 390.9 cm^{-1}) were used. After spike removal, the raw spectra were interpolated in the range from 400 to 1750 cm^{-1} to achieve an equal distribution of data points across the whole spectral range. A distance of 1.45 cm^{-1}, corresponding to the average spectral resolution in the experiment was chosen as distance between variables. Subsequently, a baseline for each spectrum was estimated by asymmetric least square smoothing[177] and subtracted from the respective spectrum, followed by vector normalization of the baseline corrected spectrum. An average spectrum was calculated for each grain/sample from the respective spectra.

The Raman maps of the plant cross sections from the data set Cucumber and Sorghum were obtained by Dr. Zszusanna Heiner and Ingrid Liedkte. A detailed description of the Raman experiments is published by Zeise *et al.* for the data set Cucumber[7] and published by Heiner *et al.* for the data set Sorghum.[152] In summary, the cross sections were measured with a spatial distancing of 1 μm using a 532 nm CW laser and a laser power of 10 mW (intensity $1.7 \cdot 10^6$ Wcm^{-2}, accumulation time, 1 s). Spikes were removed and spectra were calibrated using Matlab (MathWorks, Inc., 2014 by collaboration partner Ingrid Liedkte.
Raman experiments on root sections from *Sorghum bicolor* (L.) Moench (line BTx623) were carried out by Nerya Zexer. The sample was measured using a 63x water immersed objective and 532 nm excitation wavelength. Data are obtained in streamline mode with an acquisition time of 30 s. More details can be found in reference [40].[40]
Further pre-processing such as interpolation, AsLS-background correction, and vector normalization were applied with respect to the specific data analysis and are discussed in Chapter 9.

3.2.4 Surface enhanced Raman scattering (SERS)

The subset of 72 *Poa alpina* pollen samples from Pollen Norway I was measured by SERS during the period of the master thesis[189] and reused for a comprehensive multiblock study here in Chapter 5.
In the SERS experiments, the water-soluble components of the pollen grains were extracted and mixed with an aqueous solution of citrate-stabilized gold nanoparticles as described by Seifert *et al.*[3] 100 μl Millipore water were added to 0.2 mg of the pollen sample. After 5 minutes, the samples were centrifuged and the supernatant was pipetted off. 2 μl of this aqueous pollen extract were mixed with 20 μl citrate-stabilized gold nanoparticles obtained[195] and 2 μl of a 0.1 M sodium chloride solution were added. Subsequently, 20 μl of this mixture were transferred to a calcium fluoride slide for the SERS measurement. The SERS experiments were performed on a Raman microscope (Horiba, Bensheim, Germany) in the focal volume of a 60x water immersion objective (Olympus, Hamburg) with a laser operating at a wavelength of

785 nm and an intensity of $2.9 \cdot 10^5\ Wcm^{-2}$. Two extracts for each sample (technical replicates) were prepared and analyzed. For each extract, 1000 spectra with an accumulation time of 1 s per spectrum were collected. This procedure yielded SERS data sets containing 144000 individual spectra in total (2000 spectra per pollen sample). The spectra were frequency calibrated using a spectrum of 4-acetamidophenol. Further pre-processing included spike removal, interpolation, AsLS-baseline correction,[177] and vector normalization as described in the previous section. The 2000 spectra for each sample (obtained from different extracts) were averaged so that in total 72 average SERS spectra were analyzed in the multiblock analyses.

3.2.5 Bright-field images

For all microscopical images the pollen grains were deposit on glass slides and investigated using a light microscope (Olympus BX23, Hamburg, Germany) with either a 20x or 100x objective (Olympus, Hamburg Germany). The images were recorded with CellSens Standard Software 1.17 (Olympus, Hamburg Germany).

3.2.6 Additional plant data

Pollen Norway I

Morphological and dry weight measurements, as well as the determination chlorophyll content of parent plants, were executed by Dr. Boris Zimmermann and Dr. Murat Bağcıuğlu, resulting in an additional data block for multiblock analysis that is discussed in Chapter 5.
During the pollination stage, the height of the flowering shoots of the parent plants was determined, using the average value for three highest flowering shoots per individual plant. Furthermore, the number of flowering shoots for each individual was determined. Plant dry mass was determined at the end of seed production life stage by cutting the parent plants at ground level and drying them at 60 °C for 24 hours.
Chlorophyll a and b concentrations were measured by ultraviolet-visible absorption measurements.[196] Leaf samples from each individual were collected during the pollination life stage. Approximately 4-8 mg of a sample were transferred to microcentrifuge tubes containing 1.5 ml N, N-dimethylformamide, and kept at +4 °C for 24 hours. The extracts were measured with a UV-Vis spectrophotometer (Shimadzu 1800, Japan) using cuvettes with 1 cm path length. Chlorophyll a and b concentrations were calculated by using absorbance values at 647 nm and 664 nm according to the equations by Porra *et.al.*[197] The chlorophyll content, morphological and dry weight data were combined in a separate, fifth data block, termed

here additional plant data. The data were normalized based on data dispersion (autoscaling) before further analysis.

Sorghum

Measurements of the plant heights and the size of the 7th leaves of the eight Sorghum plants as well as elemental analysis of were conducted by Ingrid Liedtke. A data set is obtained comprising the information about the four parameters: size of the leaf (area), cell wall density, dry mass, and plant height.
Before energy dispersive X-ray spectroscopy (EDX), the leaves were charred to ash by an established prototol.[37,198] The leaves were calcined at 550 °C for 24 h and subsequently washed with 1 ml 1 M hydrochloric acid.
The elemental analysis includes the contributions of Carbon, Calcium, Magnesium, Potassium, and Sodium. Both data set are used in a multiblock analysis discussed in Section 9.2.2.

3.3 Data analysis

3.3.1 Data management and data pre-processing

Data management, spectral pre-processing, and multivariate analysis were carried out using in-built functions of the software Matlab (The Mathworks, Inc., Natick, MA, USA).
The data management includes Matlab routines that transform single spectra or spectra maps stored into ASCII-files into formatted data set taken into account the structure and purpose of the experiment. Spectral pre-processing is individually adapted for each data set including in the majority the pre-processing steps interpolation (with the exception of FTIR and additional plant data), baseline correction (with the exception of additional plant data) and a normalization.
Interpolation were applied to Raman and MALDI data using the in-built Matlab function *interp1*.
The baseline correction was carried out as proposed by Eilers.[177,178] Subsequently, the baseline was subtracted from the spectrum.
Normalization of the data were executed using vector normalization, autoscaling (for discrete data as EDX and additional plant data), and extensive multiplicative scatter correction (EMSC) for FTIR data.
Before, calculation of a EMSC model smoothing is carried out on the spectra using the algorithm proposed by Savitzky and Golay[174] using a polynomial of the second order and an

optimized windows size. To estimate the optimal windows size in the smoothing algorithm, the procedure proposed by Zimmermann *at el.*[194] was applied for each pair of the 5 pollen species of the 50 samples of the data set Pollen Norway IIa. 2-class Partial least square discriminant analysis (PLS-DA) was applied for window sizes between 3 and 23 with varying amount of latent variables. For visualization, the classification results of one pair (*Hordeum bulbosum* and *Poa alpina*) are presented in Figure 3.1. The optimal number of smoothing points would have less latent variables and the most correct classified spectra. Here, using a windows size higher than 11 and 4 or 5 latent variables, >90 % of spectra would be classified correctly. In order to avoid over-smoothing of the data the smallest number (here 11) was chosen.

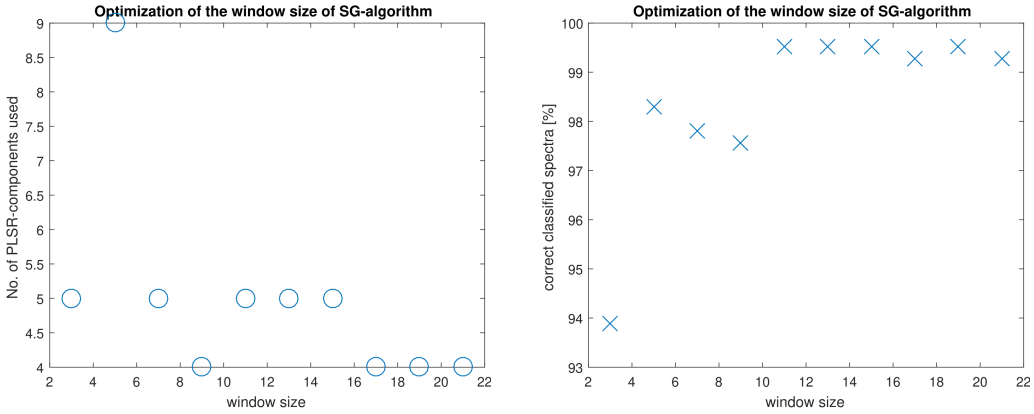

Figure 3.1: Classification results of 2-class PLS-DA from FTIR spectra of *Hordeum bulbosum* and *Poa alpina* for optimization of the windows size. **(Left)** Number of latent variables with respect to the windows size. **(Right)** Correct classified spectra with respect of the windows size.

The optimal window size was estimated for all ten pairs and the median of a windows size of 9 was applied as the parameter for smoothing of the data set. The model was carried out using linear and quadratic terms of the model.
The optimization of the windows size and the calculation and application of the model was executed using Matlab-functions developed in collaboration with Prof. Dr. Stephan Seifert and Prof. Dr. Achim Kohler.

3.3.2 Multivariate analysis

Hierarchical cluster analysis

Hierarchical cluster analysis (HCA) was applied using euclidean distances and Ward's algorithm[58] on the specific pre-processed data sets. The in-build Matlab functions *pdist* and *linkage* calculate and the function *dendrogram* visualizes the results. HCA was carried out to

i) assess the relation between different samples in the case of the data set Pollen Norway IIa (Chapter 7) and ii) to explore the heterogeneity of imaging data set of MALDI mass spectra (Chapter 8) and Raman data (Chapter 9).

Principal component analysis

Principal component analysis (PCA) was conducted using in-built Matlab routines. For evaluation of the score values, Kruskal-Wallis H-test[199] was applied, using *kruskalwallis*-function in Matlab on the classification problem. The Kruskal-Wallis H-test as a non-parametric statistical test was chosen after assessment of the data sets regarding their normal distribution. Some data sets, e.g. obtained from SERS experiments do often not show normal distribution.[3] The test is used to prove the null hypothesis that the distribution of the data within each respective classification group is equal. A p-value below 0.05 indicates a significant difference in this distribution for at least one of the groups.
Furthermore, d-values were estimated for evaluation of the PCA models using one-way multivariate analysis of variance (MANOVA) in combination with Bartlett test[200] on the PCA score values of the first ten PCs (covering at least 90 % of the explained variances in each PCA model) obtained by the *manova1*-function in Matlab. The d-values estimate the dimensionality of a group based on the non-random variances in the group means by giving a value between 0 and N-1, with N being the amount of groups that need to be identified.
The combination of PCA and statistical tests is executed: I) On the data of the sample set Pollen Norway I and discussed in Chapter 4 and 5.
II) as a chemometric tool for exploration is in addition carried out on the data sets Pollen Israel (Chapter 6), Pollen Norway II (Chapter 7)
III) On the Raman mapping data discussed in Chapter 9.

Consensus principal component analysis

Using Matlab routines developed in colllaboration with Prof. Dr. Achim Kohler, consensus principal component analysis (CPCA)[62,63] was used to combine data of different individual methods. In order to apply CPCA, all the data sets need to have the same sample dimension, and the order of the samples should be identical for all data sets included in the analysis. Therefore, some data have to be averaged before further analysis. For the common representation of the loadings of all the different kinds of data in one correlation loading plot as result, thresholds were defined for the respective data types and those positive/negative peaks above/below the respective threshold are represented in the unified plot. For clarity, all other spectral variables are not shown in these plots. Only the variables belonging to discrete data blocks (additional plant data and EDX) are displayed as a whole. CPCA is here used for:

I) The assessment of the subspecies variation in grass pollen (Pollen Norway I) and their comprehensive characterization using a combination of FTIR spectroscopy of homogenized pollen grains, Raman spectroscopy, SERS of the water-soluble extract, MALD-TOF MS of the acid extracts, and additional plant-related information (Chapter 5).
II) In Section 6.2, the evaluation of Raman data in combination with MALD-TOF MS data using pollen from wild-type and*SbLsil*-mutants of *Sorghum bicolor* (L.) Moench (line BTX-623), as well as different growth conditions (Pollen Israel).
III) The combination of two different spectral ranges, when a part of the spectrum is omitted. CPCA is executed here on Pollen Norway IIa and discussed in Section 7.2.2.
IV) The assessment of the variation between five different grass pollen species (Pollen Norway IIa) and their comprehensive characterization using a combination of FTIR microspectroscopy of embedded single pollen grains, Raman microspectroscopy, and MALD-TOF MS of the acid extracts (Section 7.4).
V) the combination of Raman microspectra from different substructures extracted from Raman maps, and additional plant-related information in order to explore the differences between cross sections from wild-type and*SbLsil*-mutants of *Sorghum bicolor* (L.) Moench.

Partial least square discriminant analysis

For PLS-DA and other supervised multivariate methods, a corresponding target matrix towards the data matrix needs to be built. This target matrix stores the assignment of one spectrum to a specific class with 1 (assigned to the class) and 0 (not assigned to the class). A PLS-DA model is usually trained with several latent variables. Using the *plsregress*-function in Matlab to optimal amount of latent variables can be estimated using 10-fold cross validation. Two different types of external validation of the trained model were applied in this thesis.
The first one is the leave-one-out cross validation (full-CV), where each model was trained using the whole data set except one spectrum. The classification of the remaining spectrum is calculated and a new model is trained using the whole data set except one other spectrum. This procedure is repeated for all spectra of the data set.
In the other approach, the classification analyses were conducted by splitting the spectral data set into two parts (here 50 % :50 %) where one part is the training set and the other part is used for validation the model. The data in these two sets can be either randomly chosen from the whole data set or systematic according to the hierarchy (e.g. leave-one-sample-out, leave-one-population-out). .
PLS-DA was conducted here to:
I) Assess the variation within a hierarchically structured framework of grass pollen, comprising variation between species, populations, and growth conditions (Pollen Norway I discussed in Chapter 4).

II) Compare different pre-processing approaches and their performances in model-based classification of the data set Pollen Norway IIa and Pollen Norway IIb (discussed in Chapter 7).
III) Classify MALDI imaging mass spectra from pollen mixtures from the data set Pollen Berlin (discussed in Chapter 8).

Non-negative matrix factorization

Non-negative matrix factorization was conducted on the FTIR data from Pollen Norway IIa discussed in Chapter 7 and on the MALDI MS imaging data of the data set Pollen Berlin discussed in Chapter 8. Decomposition into components and the corresponding relative contributions of the respective spectra was achieved by using the *nnmf* function in Matlab and the alternating least square (ALS) algorithm.[201] The optimized number for decomposition was evaluated by eye, respectively.
The factorization of the data set Pollen Norway 2 was carried out using six components. Subsequently, the components were sorted by eye and reconstructed by subtraction of the components from the data matrix. More details can be found in Section 7.2.3.
In the case of the imaging data set from Pollen Berlin, NMF was applied using four components. For visualization of the results, the relative contributions for each spectrum of the MALDI map were min/max normalized so that the image can be presented as four classification images.

Machine learning approaches: artificial neural networks and random forest

Artificial neural networks (ANN) and random forest (RF) were applied in order to classify the samples of the data set Pollen Norway IIa (Chapter 7). In addition, the MALDI imaging mass spectra of pollen mixtures (Pollen Berlin, Chapter 8) were evaluated using ANN and RF. The data set Pollen Norway IIa were split into training and test set. In the case of the MALDI image, references spectra form the training set and the models wer tested on the imaging data. Feed-forward ANN[202] were applied to the data with a number of input neurons that correspond to the amount of variables in the spectra, 50 neurons in the hidden layer, and an amount of output neurons that corresponds to the number of the classes given by the target matrix. To employ ANN, the Matlab functions *patternnet* and *train* with 70 % of the training set were used for training, 25 % for validation and 5 % for internal testing, followed by validation of the independent test set.
Random forest classification was applied by using the *treebagger* function in Matlab. External validation of the test set was executed using the *predict*-function.

4 Classification of pollen in a hierarchical framework of variances using MALDI TOF MS

Parts of the results presented in this chapter are published in: Diehn, S., Zimmermann, B., Bağcıuğlu, M., Seifert S., Kohler A., Ohlson M., Fjellheim S, Weidner S., and Kneipp J.. Matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS) shows adaptation of grass pollen composition.*Sci Rep* **8**, 16591 (2018). (https://doi.org/10.1038/s41598-018-34800-1).

Data sets obtained from pollen, and other complex biological systems are often organized in a hierarchical framework containing the phylogenetic relationships between samples, as well as other conditions, e.g., different populations, different treatments during plant growth, or genetic background.[15,16] As a consequence, spectra from the biological samples can provide information about these different sources of variation, such as species, populations, or growth conditions. Matrix-assisted laser desorption/ionization time of flight mass spectrometry (MALDI TOF MS) has shown to provide a specific fingerprint of pollen which allows discrimination between different pollen genera and species.[1,2,107]
In this chapter, a data set of MALDI TOF mass spectra from grass pollen of the same family (Poaceae) is explored with respect to a hierarchical framework based on species, sub-species level, and different growth conditions (Pollen Norway I, see Chapter 3). This framework is presented in the scheme in Figure 4.1. The well-designed and characterized data set of 272 mass spectra is analyzed focusing on variations induced by species, population, and growth conditions.

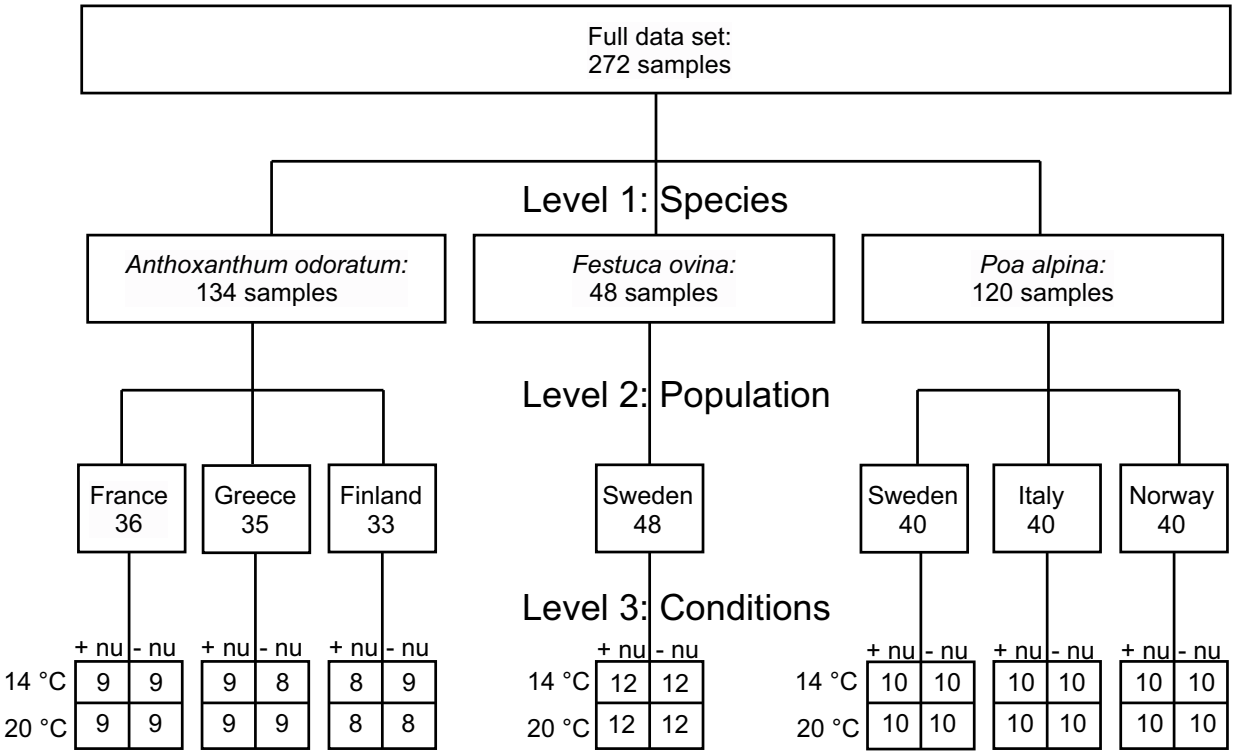

Figure 4.1: Scematic representation of the 272 mass spectra from the sample set Pollen Norway I presented as a hierarchical framework. The data set comprises spectra from three different species, seven different population and four different growth conditions. The numbers in the squares of the third level correspond to the amount of individual genotypes in each population that are cloned four times in order to investigate one clone for each growth condition.

4.1 Exploratory analysis of a hierarchically structured data set

First, the averaged spectra for each grass species are compared. Subsequently, chemometric methods, namely partial least square discriminant analysis (PLS-DA) and principal component analysis (PCA) are applied in order to explore the variances in the data set.

In Figure 4.2, the averaged spectra for each of the three pollen species (*Anthoxanthum odoratum*, *Festuca ovina*, and *Poa alpina*) are shown. The spectra present fingerprint-like patterns in the chosen mass range. The peaks can be associated with unidentified oligosaccharide and other biopolymers.[1,2,107]

The mass spectra of *Anthoxanthum odoratum* pollen grains contain particularly three strong peaks at *m/z* 5324, 5748, and *m/z* 6332, while in the case of *Festuca ovina* and *Poa alpina* more peaks are present. Between the mass range *m/z* 5450 and 5800, the mass spectra of *Festuca ovina* pollen have five characteristic peaks, while in *Poa alpina* only one dominant peak at *m/z* 5718 appears. Comparing the mass range from *m/z* 5850 to 6200 in the mass spectra of

both species, only two peaks at *m/z* 6004 and 6068 show up in the mass spectra of *Festuca ovina* pollen, while more peaks can be detected in the ones of *Poa alpina*. In particular, the peak at *m/z* 6128 is prominent in the *Poa alpina* mass spectra. After *m/z* 6000 the amount of peaks in all three spectra is limited. Spectra of *Anthoxanthum odoratum* pollen grains are not showing dominant peaks in this mass range. In the case of *Poa alpina* mass spectra, two peaks (at *m/z* 6740 and 6782), as well as several small peaks around *m/z* 8110 contribute to the species-specific pattern. In the mass spectra of *Festuca ovina* pollen a remarkable peak at *m/z* 7232 and two peaks at *m/z* 8024 and *m/z* 8056 occur in the upper range.
In comparison to vibrational spectroscopy, where a similar chemical composition of the pollen grains may lead to spectra that look nearly identical, MALDI-TOF MS provides highly species-specific patterns for the pollen of the three grass species, which enables accurate identification of the three pollen species by eye.

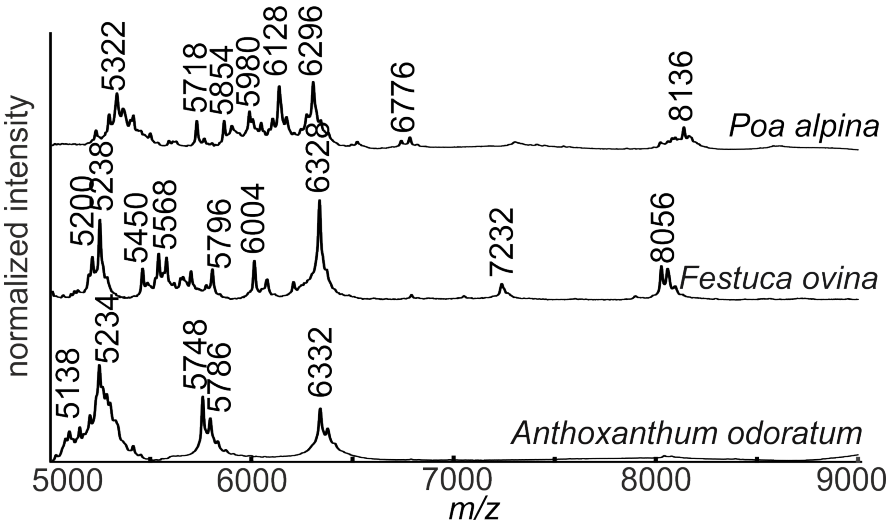

Figure 4.2: Pre-processed and averaged mass spectra in the range from *m/z* 5000-9000 for the three grass species *Anthoxanthum odoratum*, *Festuca ovina* and *Poa alpina*.

Since it has been shown that the mass spectra of pollen are species-specific, as a second step the averaged spectra for the different population of the species *Poa alpina* will be investigated(Figure 4.3).
In general, the averaged spectra of different populations show more similarities, but some differences can still be detected by eye. In particular, the peak at *m/z* 5854 is prominent in the spectra of the population Norway but weak in the other two populations. In general, the pattern between *m/z* 5900 and 6300 shows differences, regarding the different populations.

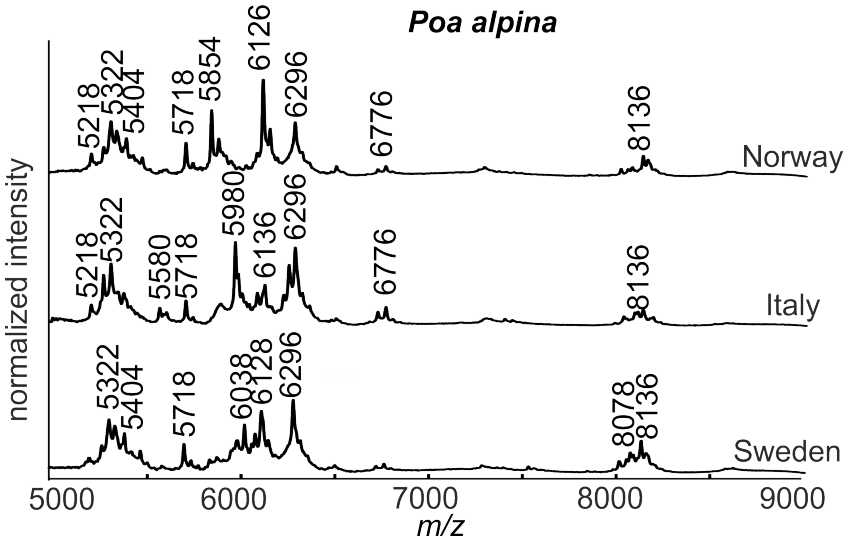

Figure 4.3: Pre-processed and averaged mass spectra in the range from *m/z* 5000-9000 for the three different populations Sweden, Italy and Norway from the grass species *Poa alpina*.

It can be concluded that the MALDI TOF mass spectra show a specific peak pattern depending on both the species and population. Even so, the peak patterns are based on averaged spectra. An exploration of all spectra would lead to a more accurate characterization.
The most common approach for such an explorative analysis is PCA, that provides a first view of the source of variances within a data set, due to the weighting of the variables of the spectra. A common visualization of the PCA results is a 2D scatter plot of the scores from two chosen components and the corresponding loadings. Usually, more than two component show interesting features, leading to time-consuming evaluation and interpretation of scatter plots and loadings. Furthermore, effects that cause higher variances in hierarchically structured data sets may overlay the grouping or discrimination of a smaller source of variation within the same data set.
Figure 4.4 shows the scores plot of the PCA for the same data set, colored according to the three different pollen species (Figure 4.4, left) and the seven different populations (Figure 4.4, right). As indicated by the scores plot on the left, the three species can be differentiated by the first PC (Figure 4.4, left), while the seven populations cannot be discriminated (Figure 4.4, right). Nevertheless, further evaluation of the PCA results can give insights into the variances beyond the 2D- represention of the scores plot. Further statistical approaches can be applied to the score values in order to simplify the evaluation of the PCA.

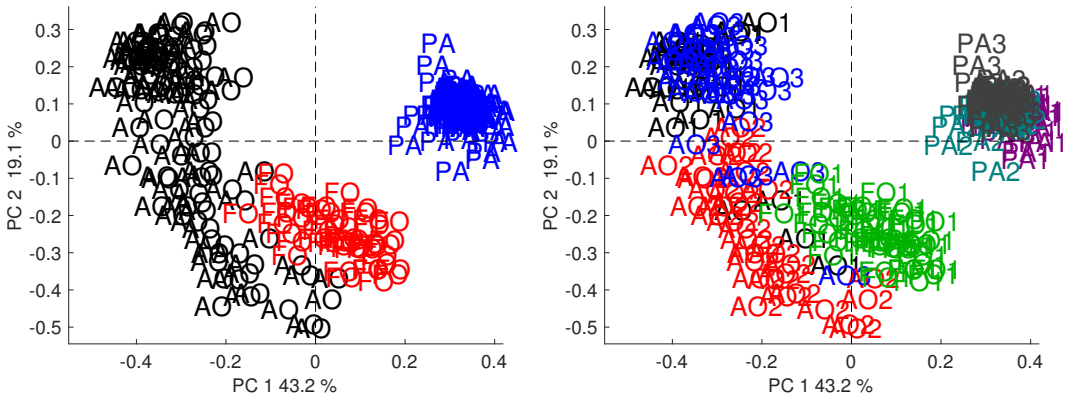

Figure 4.4: (Left) Scores plot of the first and second PC of the 272 mass spectra with a coloring with respect to the three different species, *Anthoxanthum odoratum*, black, *Festuca ovina*, red and *Poa alpina*, blue. **(Right)** Scores plot of the first and second PC of the 272 mass spectra with a coloring with respect to the seven different population.

4.1.1 Distribution of the score values

A PCA decomposes the data set into scores and loadings (see Chapter 2.3.2) and sorts the components on the base of their explained variances in decreasing order. As shown above, the visualization of the score values is colored with respect to the classification problem. In the data set, there can be respectively three or seven colors depending on species and populations. The number increases for the combination of population and growth conditions up to 28 different groups and therefore 28 colors are possible for the scores. This would lead to an overlap of several colors.

In an explorative analysis, the distributions of the score values of each PC related to these groups are relevant. Some descriptive statistical tests, such as ANOVA, enable a calculation of a p-value that corresponds to a distribution. Since the data sets can have outlier behavior due to biological or technical variances, such a behavior could also be relevant in a comprehensive description of the data. In this respect, a non-parametrical test has been chosen to evaluate the score values.

Here, the Kruskal-Wallis H-Test[199] is used to evaluate the distribution of the score values with respect to, e.g., the species, or populations. The test uses the null-hypothesis that the distributions of all groups are equal. The null-hypothesis is rejected with a p-value below 0.05. Regarding the presented data set, the obtained p-values for the first principal component (PC 1) is always below 0.05 for both the distribution of the score values of the three species and of the seven populations. The rejection of the null-hypothesis for all the cases indicates that the distribution of one of the groups differs from the remaining distributions. A low p-value cannot guarantee, that all distributions differ from each other. An evaluation of the scores using only Kruskal-Wallis H-Test is inefficient here for some applications.

Figure 4.5 shows the box plots for the distributions of the first PC regarding three (Figure 4.5,

left) and seven groups (Figure 4.5, right) for the species and populations, respectively.

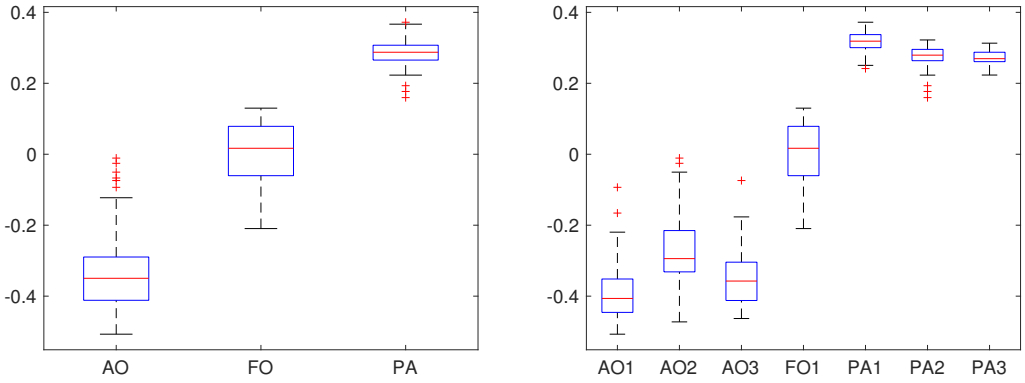

Figure 4.5: Box plots for the distributions of score values for PC 1 with respect to **(left)** three species and **(right)** seven populations. Species and populations were abbreviated to AO, *Anthoxanthum odorathum*, FO, *Festuca ovina* and PA, *Poa alpina* as well as AO1, *Anthoxanthum odorathum* France, AO2, *Anthoxanthum odorathum* Greece, AO3, *Anthoxanthum odorathum* Finland, FO1, *Festuca ovina*, PA1, *Poa alpina* Sweden, PA2, *Poa alpina* Italy, PA3, *Poa alpina* Norway.

The box plot of the PCA from the 272 mass spectra with respect to the three different species (Figure 4.5, left) confirms that the distributions of the scores (Figure 4.5, left, blue boxes) as well as the median of each distribution (Figure 4.5, red lines) are not equal.
In contrast, the box plots of the score distribution with respect to the different populations (Figure 4.5, right) show fewer differences, especially for the three populations of *Poa alpina*. The Kruskal-Wallis H-Test gives low p-values in the case that just one of the groups differs greatly from the others, reducing the information of a score plot into one value that indicates if a discrimination of groups is more or less significant. To estimate how many of the defined groups can be separated, MANOVA and Bartlett[200] test can be applied.

4.1.2 Dimensionality of the score values

MANOVA is the multivariate alternative for ANOVA and with the Matlab function *manova1* the variation within multivariate means of groups can be assessed. Here, the Bartlett test, included in the statistics, can be applied to the score values of more then one PC. As a result, information about the dimensionality of the reduced data can be obtained. The null-hypothesis proves if the group means are equal. If it is rejected, a new null-hypothesis if the means are linear depended from each other is tested. A calculated d-value indicates the dimensionality of the data, which is N-1 as maximum.
As an example, the d-values for the hierarchical data set with respect to the three different species is 2, while with respect to the seven different population is 6. This indicates that the species as well as the populations, can be discriminated using the first 10 components, that is

explained 90 % of the total variance.
The underlying Bartlett test is calculated based on the multivariate distances of each of the groups. The distances can be visualized using the *manovacluster*-function in Matlab. Similar to a dendrogram based on hierarchical cluster analysis, *Ward's*-algorithm[58] is used to form clusters with respect to the defined groups. Figure 4.6 shows the resulting dendrograms. In addition to the d-values the corresponding dendrograms indicate which groups have lower distances to each other. In Figure 4.6 (right), the dendrogram shows the clustering of the scores from the seven different populations. Each of the three clusters contains the data of the populations belonging to just one species.

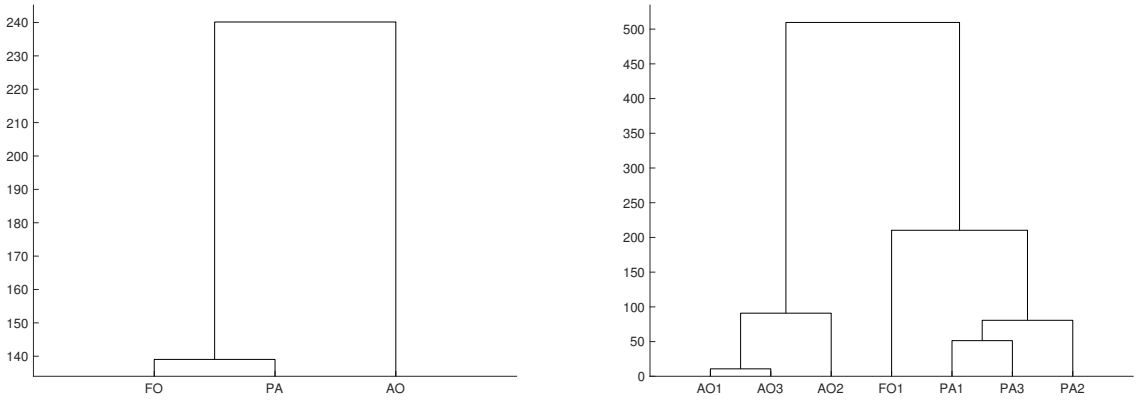

Figure 4.6: (Left) Dendrogram obtained after MANOVA of the score values from PC 1-PC 10 with respect to three species. **(Right)** Dendrogram of the score values from PC1-PC10 with respect to seven populations. Species and populations were abbreviated to AO, *Anthoxanthum odorathum*, FO, *Festuca ovina*, and PA, *Poa alpina* as well as AO1, *Anthoxanthum odorathum* France, AO2, *Anthoxanthum odorathum* Greece, AO3, *Anthoxanthum odorathum* Finland, FO1, *Festuca ovina*, PA1, *Poa alpina* Sweden, PA2, *Poa alpina* Italy, PA3, *Poa alpina* Norway.

To summarize, the combination of one p and one d-value is an extension to the explorative analysis using PCA. Statistical tests on the reduced data set give insight into the distribution of the score values and, complementary, the dimensionality of the data with respect to the biological question. Using a hierarchical data set structure including the species and populations, it can be concluded, that a predictive analysis regarding the classification of the different species and population is possible.

4.2 Discrimination of pollen spectra containing variation of species, populations and growth condition

Since the explorative analysis indicates a successful differentiation of the mass spectra from different species and populations, a model can be trained to classify the spectra with respect

to the biological question of interest. PLS-DA is commonly applied to classify spectra based on an underlying pattern. In short, PLS is an iterative algorithm that projects the spectra in a data matrix X onto either a vector Y (PLS1) or a target matrix Y (PLS2). In terms of classification, a target matrix based on the respective affiliated species or population is filled with either 0 or 1 (for more details see Chapter 2.3.3).

4.2.1 Optimization of a PLS-DA model

Since *Poa alpina, Festuca ovina*, and *Anthoxanthum odoratum* belong to the same grass family,[26,27] and their pollen grains have similar morphology. This makes the discrimination by the light microscope impractical. Zimmermann *et al.* discussed that the chemical fingerprint obtained by FTIR enables a robust classification for the three grass pollen species.[15] Other studies show the potential of MALDI-TOF MS of separating orders and species, based on the mass spectra of the pollen.[1,2] PLS models can be calculated in order to give insight into how specific the peak pattern is for a particular for a particular species.
Before training a PLS model, an appropriate amount of components (latent variables) has to be estimated. A high amount of latent variables would lead to a model resulting in high success rates for the classification of the spectra from the training set. However, it would be less efficient with an independent test set. The optimal amount of latent variables can be estimated using the error of the prediction, usually the root mean square error (RMSE) as a function of the amount of components.
For the validation of the model, the data set is randomly split into a specific amount of subsets, called folds. One fold is left out of the training set and used for testing. After the training, the RMSE is calculated based on the classification results of the remaining fold. Permutatively, this is repeated for all folds, and the total RMSE is obtained as the average of the individual RMSEs. This procedure is repeated for a certain amount of latent variables, in order to estimate the appropriate number for a good classification and no over-fitting. Figure 4.7, left shows the resulting plot, using all 272 spectra and 10-fold cross validation obtained by the *plsregress*-function in Matlab.

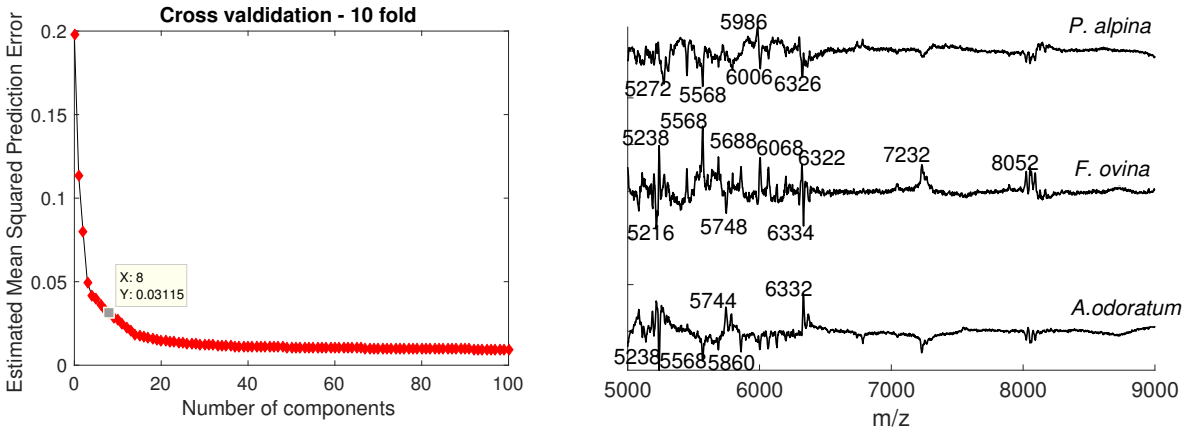

Figure 4.7: (Left) Estimation of the appropriate amount of components using the RMSE and 10-fold cross validation. **(Right)** Regression parameter for the three different pollen species *Anthoxanthum odoratum*, *Festuca ovina*, and *Poa alpina*.

The PLS-DA model is trained to discriminate between the three different species. The resulting cross-validation plot is showing the typical exponential curve. The lowest error could be achieved using at least 14 components. To avoid over-fitting, a compromise between the number of latent variables and the error leads to an amount of eight latent variables. A new PLS-Model is calculated using the eight latent variables and the leave-one-spectrum-out-approach (full-CV).
Based on the training set, the regression parameter for each of the three columns in the target matrix Y is calculated by PLS-DA (see Chapter 2 for more details). The regression parameter indicates which variables the trained classification is based on. Figure 4.7 (right) shows the regression parameters for the three species *Anthoxanthum odoratum*, *Festuca ovina*, and *Poa alpina*. Similar to the loadings received by PCA an interpretation of the regression parameter can give insights into the source of variation. In contrast to PCA, PLS-DA is a supervised classification method and consequently depended on the assignment of the spectra within the right class. The investigation of the variances within data is, therefore, more powerful using loadings from a PCA.

In the full-CV approach, the data set of 272 spectra is divided into 271 spectra for the training set and one spectrum as the test set. The regression parameters are used to calculate how well the test spectrum would be classified into one of the given classes, which means the three different species. For each of the three classes, a value from 0 to 1 is obtained as a result. The classification is based on the *winner-takes-it-all*-approach, where the classified class is the one with the highest value. This process is repeated for all 272 spectra.
Table 4.1 shows the results of the PLS-DA using eight latent variables. All 272 spectra are classified as their correct species. MALDI-TOF MS enables an accurate and fast identification of different grass pollen as an alternative for pollen forecast by light microscopy.

Table 4.1: Classification of 272 pollen spectra according to their species applying PLS-DA using eighth latent variables and full-CV.

affiliation / identified by PLS-DA as	*A. odoratum*	*F. ovina*	*P. alpina*
A. odoratum	104	0	0
F. ovina	0	48	0
P. alpina	0	0	120
Success rates	100 %	100 %	100 %

4.2.2 Classification of grass pollen spectra from different species and populations

Since the variation in the whole sample set is structured in a hierarchical framework (Figure 4.1), PLS-DA was applied also for the classification according to populations and growth conditions using the whole data set.

For the classification of the 272 spectra in the full-cross validation approach the following overall success rates were estimated: species, SR = 100 % (Table 4.1), populations, SR = 94.5 % (Table 4.2), and in addition, the populations divided into the four growth conditions, SR = 26 % (not shown). For each of the three levels, a p value $\ll 0.01$ is calculated for the discrimination of the three, seven and 28 groups.

The discrimination of different populations shows high success rates (Table 4.2). Table 4.2 indicates that all misclassified spectra belong to *Anthoxanthum odoratum*. Five of 36 spectra from population *Anthoxanthum odoratum*, France were classified as the two other populations, but still as *Anthoxanthum odoratum*. 10 of 33 spectra of the population *Anthoxanthum odoratum*, Finland are misclassified as *Anthoxanthum odoratum*, France.

Nevertheless, in comparison to previous studies using FTIR on the same pollen samples, the chemical pattern, obtained by MALDI-TOF MS is highly population-specific. The high throughput FTIR measurements were limited to 77 % success rate,[15] compared to 94.5 % in the MALDI MS results discussed here.

Table 4.2: Classification of 272 pollen spectra according to their populations using PLS-DA and eight latent variables.

affiliation / identified by PLS-DA as	*A. odoratum* France	*A. odoratum* Greece	*A. odoratum* Finland	*F. ovina*	*P. alpina* Sweden	*P. alpina* Italy	*P. alpina* Norway
A. odoratum France	31	0	10	0	0	0	0
A. odoratum Greece	2	35	0	0	0	0	0
A. odoratum Finland	3	0	23	0	0	0	0
F. ovina	0	0	0	48	0	0	0
P. alpina Sweden	0	0	0	0	40	0	0
P. alpina Italy	0	0	0	0	0	40	0
P. alpina Norway	0	0	0	0	0	0	40
Success rate	86 %	100 %	70 %	100 %	100 %	100 %	100 %
Overall Success rate	94.5 %						

Previous FTIR-studies discussed, that the species-specific differences between the grass pollen are strongly dependent on the ratios of lipids, proteins, and carbohydrates in the pollen grains.[14,15] With mass spectrometry, these large biomolecules and their respective fragments can be detected as a unique pattern with high sensitivity. Due to this large number of possible biomolecules, a detailed interpretation of the peaks requires further experiments, such as a more sensitive MS/MS approach.[1] Also without such a detailed understanding of the peak pattern, it has been demonstrated here that a fingerprint of the biochemical composition of the pollen grains can be obtained, which enables a classification between species from the same family and in particular between different populations.

4.2.3 Variation of pollen spectra regarding different growth conditions and genotypes

Table 4.4 contains the success rates and p-values for the different environmental effects on the pollen grains. A p- value below 0.05 rejects the null- hypothesis that all groups have the same distribution on a significance level of 5 %. The p-values were determined by the PCA score values of the respective data set.
Using PCA, the spectra of some populations can be discriminated according to the defined

design factor ($p < 0.05$). *Poa alpina* shows a significant influence of the growth condition in the mass spectra. In contrast, the success rates for the identification for the other spectra are quite low. Only the identification of the *Poa alpina* mass spectra for the growth conditions and in particular the different nutrient conditions in the population Italy can be classified with high success rates by PLS-DA and with PCA (Table 4.4).

As discussed above, MALDI-TOF MS enables detection of oligosaccharides and other large other biomolecules, which could give a fingerprint of the nutrients composition inside the pollen grains. The chemical composition inside the pollen grains is highly species- and population-specific. Therefore, for *Poa alpina* higher success rate can be detected, but only in the population Italy a significant discrimination of different nutrient conditions is possible. Figure 4.8 shows the averaged spectra with the standard deviation for all mass spectra of pollen from individual plants of the four growing condition of the parent plants within the population *Poa alpina*, Italy.

The averaged spectra are almost identical with small exceptions in the mass range from m/z 5400 to m/z 5900. In particular, the peak at m/z 5474 exists in the mass spectra for pollen of plants that are grown up at 20 °C but is less dominant in the mass spectra from pollen grains of plants that are grown up at 14 °C . This peak is also strongly dependent on the investigated genotype, marked as the high standard deviation of the spectra at this position.

In addition, the peak at m/z 5584 is more prominent in mass spectra from pollen of parent plants that were treated with additional nutrients. Apart from the the variability of the peak at m/z 5474, also other peaks show fluctuations between genotypes, visualized by the gray area representing standard deviation. Especially the peak at m/z 5900 varies between different genotypes in all four conditions.

Table 4.3: Success rates for classification of the pollen spectra from samples of the four growth conditions for an independent test set and for full-CV using PLS-DA, as well as p-values of the first principal component calculated using Kruskal-Wallis test.

Growth conditions	Pollen species/ population	No. of latent variables	Succes rates	p-value PC 1	p-value lowest
temperature (14 °C /20°C)	*A. odoratum* France	6	56 %	0.0665	0.0290, PC9
	A. odoratum Greece	7	74 %	0.4093	0.0017, PC7
	A. odoratum Finland	8	30 %	0.3132	-
	F. ovina	11	77 %	0.5362	-
	P. alpina Sweden	7	78 %	0.1762	0.0119, PC7
	P. alpina Italy	7	80 %	0.062	0.0149, PC5
	P. alpina Norway	7	78 %	0.0173	-
nutrients (+nu/-nu)	*A. odoratum* France	7	47 %	0.4477	-
	A. odoratum Greece	7	46 %	0.2761	0.0175, PC5
	A. odoratum Finland	7	48 %	0.1303	-
	F. ovina	12	46 %	0.6501	-
	P. alpina Sweden	8	65 %	0.8070	0.0284, PC10
	P. alpina Italy	8	98 %	0.0017	
	P. alpina Norway	8	83 %	0.1046	0.0038, PC5

Table 4.4: Success rates for classification of the pollen spectra from samples of the four growth conditions for an independent test set and for full-CV using PLS-DA, as well as p-values of the first principal component calculated using Kruskal-Wallis test.

Growth conditions	Pollen species/ population	No. of latent variables	Succes rates	p-value PC 1	p-value lowest
all conditions	*A. odoratum* France	8	42 %	0.0226	-
	A. odoratum Greece	7	34 %	0.5733	0.0485, PC5
	A. odoratum Finland	6	6 %	0.13	0.0544, PC3
	F. ovina	11	33 %	0.8928	0.0446, PC9
	P. alpina Sweden	9	35 %	0.5412	0.0294, PC7
	P. alpina Italy	7	63 %	0.0038	
	P. alpina Norway	7	70 %	0.0047	

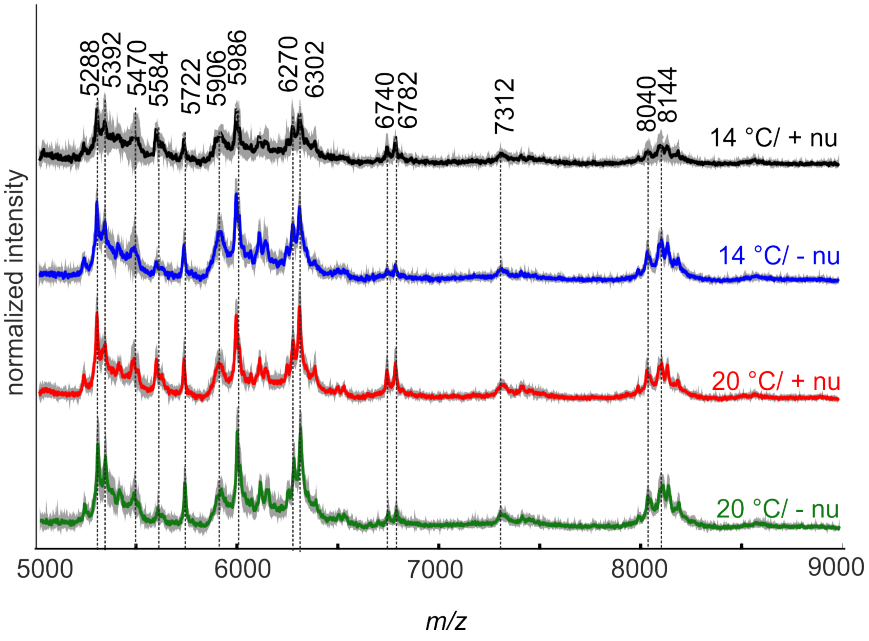

Figure 4.8: Pre-processed and averaged spectra (ten spectra each) of the population *Poa alpina*, Italy for each of the four growth conditions with their standard deviation in gray.

To investigate the influence of the growth conditions in more detail, PCA is applied to the mass spectra of all samples from population *Poa alpina*, Italy.

Figure 4.9 shows the scores plot and the loadings for the third and fifth principal component (PC).

The first and second PC are not be discussed in detail since the obtained variance is influenced

by an unknown diversity of the samples that is not related to growth conditions. The variance according to different nutrient conditions is represented by the third PC. In particular, most of the mass spectra of pollen obtained from plants that were growing without additional nutrients (Figure 4.9, blue and green) have positive values for the third PC, while most of the mass spectra of pollen obtained from plants that were growing at 20 °C and additional nutrients (Figure 4.9, red) have negative values.

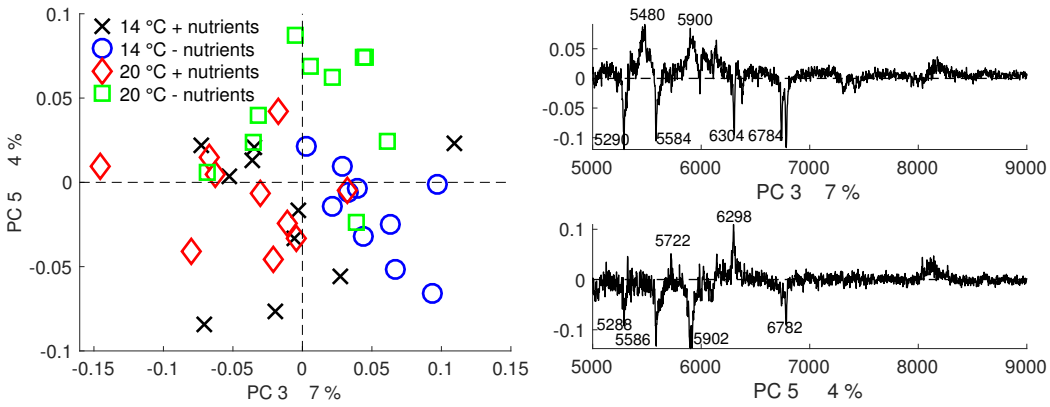

Figure 4.9: (Left) Scores plot and **(right)** loadings of the third and fifth PC using the 40 mass spectra of *Poa alpina*, Italy with a coloring with respect to the four different growth conditions, 14 °C and additional nutrients, black, 14 °C and additional nutrients, blue 20 °C and additional nutrients, red and 20 °C and additional nutrients, green.

The third PC explains 7 % of total variance and it is strongly influenced by the signals at *m/z* 5290, 5480, 5584, 5900, 6304, and the two signals around *m/z* 6780 (Figure 4.9, right). These signals have negative loading values for the third PC, meaning that these peaks are less dominant in the mass spectra of pollen of plants that were growing at 14 °C and without additional nutrients (Figure 4.9, left, blue). In comparison, these peaks are prominent in the mass spectra of pollen from plants that were growing at 20 °C and additional nutrients (Figure 4.9, left red).

In addition, most of the mass spectra of pollen obtained from plants that were growing at 14 °C and additional nutrients (Figure 4.9, left, black) have score values of the third and fifth PC that are similar to those of pollen from plants that were growing at 20 °C and additional nutrients (Figure 4.9, left, red).

Furthermore, mass spectra of pollen of plants that were growing at 20 °C and without additional nutrients (Figure 4.9, left, green) have mostly positive score values for the third and fifth PC and can be distinguished from mass spectra of pollen obtained from plants with additional nutrients (Figure 4.9, black and red) as well as pollen obtained from plants that were growing at 14 °C and without additional nutrients (Figure 4.9, left, blue) by the fifth PC. PC 5 represents 4 % of the entire variance containing signals at *m/z* 5288, 5586, 5900, 6298, and 6782.

Since the p-values and success rates for identifying the growth conditions are very low for most of the populations, other effects seem to influence the data as well. In each population, there are always four biological replicates with the same genotype treated under different conditions (Figure 4.1.

Table 4.5 shows the d-values for the effect of different genotypes within the data set. The d-values were computed by using PCA score values from the first to tenth PC. The higher the d-value the more groups can be estimated. The low values for the three *Poa alpina* populations confirm the dominant variation within the growth conditions of *Poa alpina*, Italy discussed above (Table 4.4, last section). No grouping with respect to the genotypes is possible. In contrast, the higher d-values for *Anthoxanthum odoratum* populations indicate an influence of the genotype.

These variances based on genotype compete with the variances between different growth conditions, which leads to low success rates in Table 4.5. *Festuca ovina* has the highest d-value. Nine out of twelve groups can be discriminated, as indicated by the d-value of 8. For this population the success rate success rates for classification according to each respective growth condition were low with 33 % (Table 4.4.

Table 4.5: d-values obtained using MANOVA on the score values of PC 1 to PC 10 on the data set of each population and number of theoretically possible groups (N-1) that could form based on the number of genotypes N that are present in the data of each population, assuming genotype-based cluster formation. The d-values indicate the dimensionality of the multivariate vector of the groups for each data set.

pollen species population	d- values	N-1
A. odoratum France	4	8
A. odoratum Greece	5	8
A. odoratum Finland	5	8
F. ovina	8	11
P. alpina Sweden	1	9
P. alpina Italy	0	9
P. alpina Norway	0	9

The d-values together with the success rates discussed above reflect the ability of pollen to adapt to the environmental conditions. *Poa alpina* pollen has a high phenotypic plasticity, so

larger variances between different growth conditions could be detected, while no remarkable grouping of genotypes is possible. *Anthoxanthum odoratum* and in particular *Festuca ovina* pollen show a different behavior regarding the chemical composition, which is not influenced by the different growth conditions.
The phenotypic plasticity or rigidity is correlated to the different distribution areas of the respective grass species. According to Zimmermann *et al.*, the phenotypic plasticity of *Poa alpina* depends on the restricted distribution in the alpine and other cold regions, whereas *Anthoxanthum odoratum* and *Festuca ovina* have a widespread climate distribution.[15]

To conclude, the sample set Pollen Norway I, that includes 272 pollen samples can be structured in a hierarchical framework (Figure 4.1) of three closely related grass species *Anthoxanthum odoratum*, *Festuca ovina*, and *Poa alpina*, seven different populations and four growth conditions each. PCA on MALDI mass spectra reveals a discrimination of the three pollen species. The score values of the mass spectra are evaluated by additional statistic tools, namely the Kruskal-Wallis[199] H-test and the Bartlett test.[200] As a result, the PCA can be described by two complementary values. The p-value correspond to the distribution of each group of spectra within a principal component, whereas the d-value gives insights into the dimensionality of the score values. The p-values and d-values indicate that the three species, as well as the seven populations, can be discriminated by PCA.
In addition, pollen spectra can be correctly classified in their affiliated species/populations by PLS-DA models. Variation in growth conditions was assessed by analyzing each population separately. The results give insights into the phenotypic plasticity or rigidity regarding the investigated growth conditions temperature and nutrient additions. The success rates for the classification of the variation in growth conditions differ greatly between the seven populations.
The results have shown, that MALDI MS is a suitable method to evaluate variation structured in a hierarchical framework. In particular, variation in populations can be studied. Therefore, the high sensitivity of MS data can be combined with related biological information obtained by FTIR and Raman spectroscopy, which may lead to further insights into the biochemical composition of pollen. These aspects will be discussed in Chapter 5, Chapter 6 and Chapter 7.

5 Characterization of variances in pollen spectra using PCA and CPCA

This chapter is based on the publication: Diehn, S., Zimmermann, B., Tafintseva, V., Seifert S., Bağcıuğlu, M., Ohlson M., Weidner S., Fjellheim S., Kohler A., and Kneipp J. *Front Plant Sci.* **10**, 1788 (2019).
https://doi.org/10.3389/fpls.2019.01788
The data presented here was also part of other projects; The Raman, surface enhanced Raman scattering (SERS) and matrix-assisted laser desorption/ionisation (MALDI) mass spectra were obtained in 2016 as part of a master thesis,[189] the FTIR and additional plant data was provided by collaboration partners Boris Zimmermann and Murat Bağcıuğlu.

Since no stand-alone method provides an optimal outcome in both, classification and characterization, Consensus principal component analysis (CPCA) is applied to the pollen data obtained by four complementary methods FTIR spectroscopy, Raman microspectroscopy, SERS, and MALDI-TOF MS. CPCA results were compared the results to those of PCA of each of the single data blocks. In addition, other phenotypic data on the parent plants were available and included in the analysis.
The sample set comprises pollen from one grass species, *Poa alpina* from the sample set Pollen Norway I (Chapter 3) on a sub-species level as well as different growth conditions. The parent plants originate from three different populations, within which four different growth conditions were applied to individuals of identical genetic constitution (Figure 5.1). The design of this experiment generates two separate biological questions. The first is regarding the different chemical composition of pollen from different populations in the same species. The second question relates to the differences in pollen composition as a result of different growth conditions of genetically identical plants within one population.
The results of PCA and CPCA are analyzed by statistical tests and compared for the different spectroscopic methods and separately for the different design factors, that is, population and growth condition. One of the aims is to assess the sensitivity of the multimodal characterization towards an influence of population and environmental conditions, respectively, on pollen chemistry, regardless of the hierarchical structure of the variation introduced in the specific sample set.

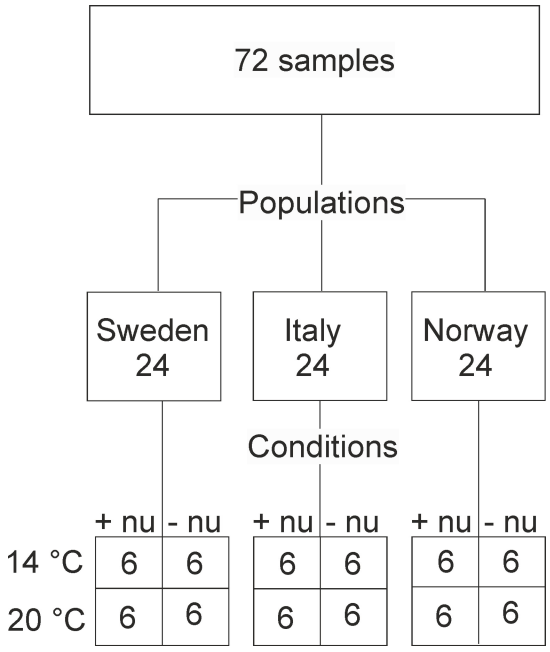

Figure 5.1: Schematic presentation of the numbers of samples (corresponding to the amount of analyzed spectra) for three populations (Sweden, Italy, Norway) and four different growth conditions (14 °C and additional nutrients, 14 °C without additional nutrients, 20 °C and additional nutrients, 20 °C and without additional nutrients). Abbreviations: +nu, additional nutrients, -nu, no additional nutrients.

5.1 Variances in pollen spectra assessed with PCA

Four different types of spectra (FTIR, Raman, MALDI, and SERS) were obtained from the 72 pollen samples, constituting four separate data blocks with 72 spectra each.

The signals in the FTIR spectra (Figure 5.2) can mainly be assigned to proteins, represented, e.g., by the amide I and amide II bands at 1669 and 1540 cm^{-1}, respectively, to lipids, exemplified by vibrations at 1156, 1467, and 1744 cm^{-1} and to sporopollenin, e.g., at 835, 1512, and 1624 cm^{-1}, in agreement with spectra reported in literature.[10,15]

The average Raman spectra in Figure 5.3 are similar to each other, albeit at slightly varying Raman shifts for some bands, suggesting small differences in the chemical composition of pollen from different populations. The bands at 1008, 1161, and 1528 cm^{-1} can be assigned to carotenoids,[31] while the signals at 526, 549, 725, 855, 1271, 1457, and 1662 cm^{-1} are due to vibrations of proteins.[13,16] The bands at 483, 1082, and 1322 cm^{-1} are assigned to carbohydrates[13,203] that can occur at high local concentrations in the pollen grains as starch deposits. Due to superposition of several molecular vibrations, some bands in the Raman spectra of pollen can be assigned to other origins as well. As examples, the bands at 1161, 1271, 1313, and 1608 cm^{-1} could also be assigned to the ferulic acid and coumaric acid building blocks in sporopollenin.[10,21] Furthermore, the band at 1608 cm^{-1} has also been associated with

mitochondrial activity.[135,204]

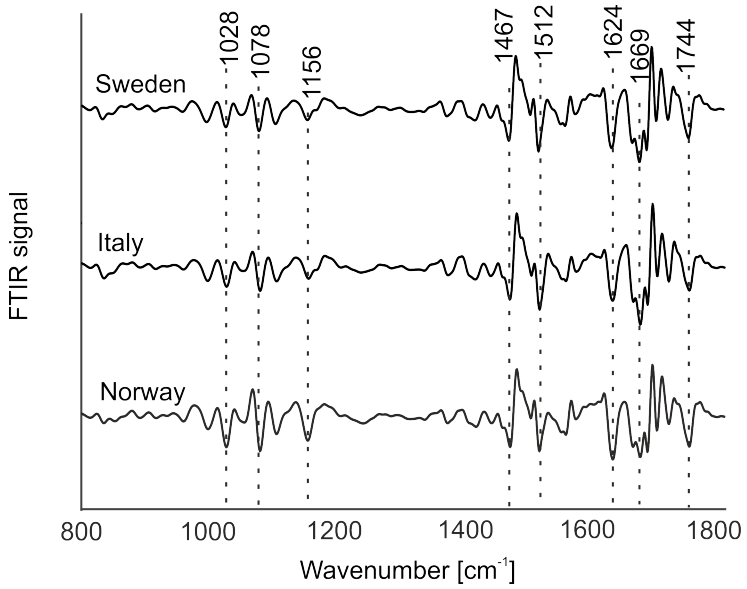

Figure 5.2: FTIR spectra of pollen from the populations Sweden, Italy, and Norway. Spectra were pre-processed by applying extended multiplicative scattering correction (EMSC) on the second derivatives and are averaged from the respective population, including samples obtained for all growth conditions. The spectra are stacked for clarity.

The SERS experiments probe the water-soluble fraction of the pollen grains due to the sample preparation as aqueous extract and the use of aqueous nano-particle solutions. Because of the high variation in the SERS spectra caused by the specifics of the SERS experiment, high numbers of spectra are needed for a reliable statistical analysis.[3] Therefore, 2000 spectra were measured from each sample, resulting in reproducible average spectra that are based on 24000 individual spectra per population. They are shown in Figure 5.4. The average spectra show characteristic bands that can mainly be assigned to vibrational modes of nucleobases, e.g., at 494, 649, 735, 802, 921 cm^{-1}[3] and amino acids, at 995, 1021, 1221 cm^{-1},[205,206] in agreement with the probing of water-soluble biomolecules extracted from the pollen.

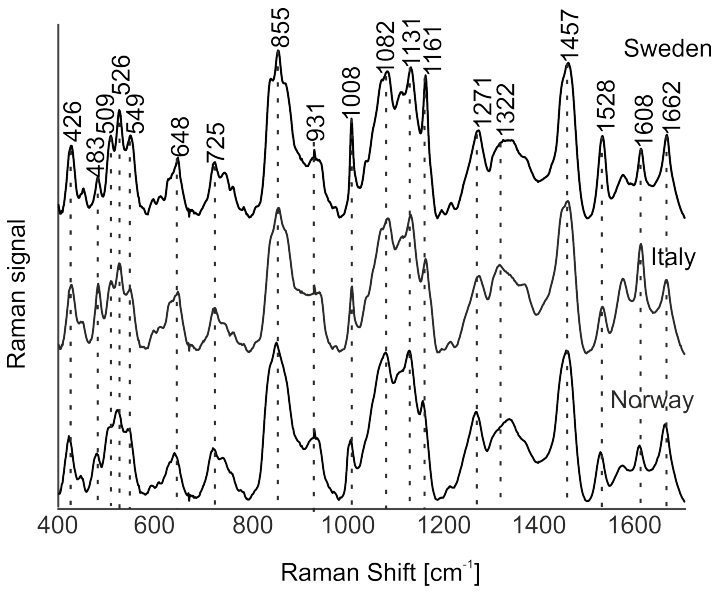

Figure 5.3: Raman spectra of pollen from the populations Sweden, Italy, and Norway. Spectra were pre-processed using Asymmetric least square (AsLS) background correction and vector normalization and are averaged from the respective population, including samples obtained for all growth conditions. The spectra are stacked for clarity.

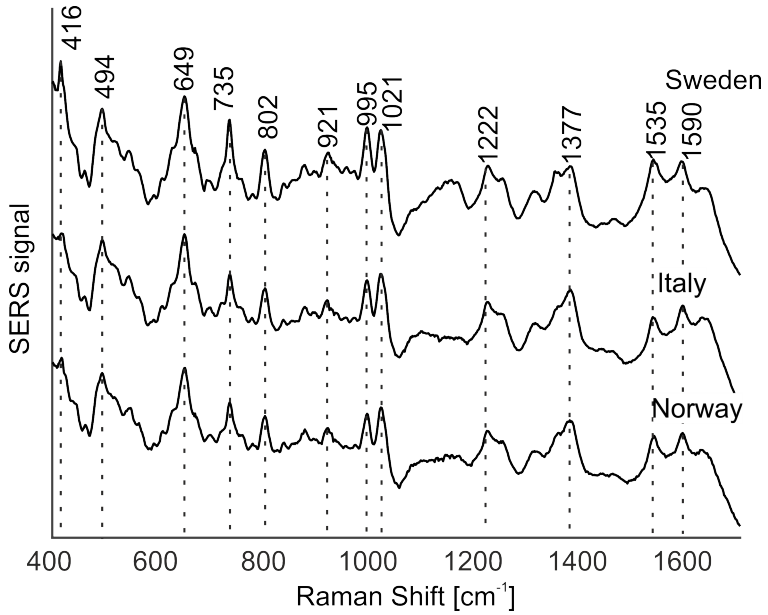

Figure 5.4: SERS spectra of pollen from the populations Sweden, Italy, and Norway. Spectra were pre-processed by AsLS background correction and vector normalization and are averaged and are averages from the respective population, including samples obtained for all growth conditions. The spectra are stacked for clarity.

MALDI TOF mass spectrometry was utilized to detect large molecules with a mass over 5 kDa. The spectral differences of pollen from *Poa alpina* are already discussed in Figure 4.3 in chapter 4. In contrast to spectra obtained by vibrational spectroscopy, in MALDI mass spectra the differences in the population averages are obvious and indicate that the pollen samples

differ in their composition in each population. The assignments of the specific peaks are not fully elucidated yet, but are most likely assigned to oligosaccharides[1,2] and larger peptides.

By PCA of the respective type of spectra/data, the pollen samples of the three different populations can be discriminated using each of the individual data blocks. Figure 5.5 shows the corresponding box plots with minimum and maximum score values to visualize the distribution of the score values of PC 1 for each population. Outliers are mainly observed for the SERS data (Figure 5.5 (C)) due to high variation owing to the specific measurement approach.[3]
The box plots for the scores of the first PC. The p-values are below 0.05 for all five data sets obtained by Kruskal-Wallis H test(numbers not shown) indicate a separation of at least one group for all these data sets.
The box plots in the left column of Figure 5.5 show that an unequivocal separation of all three populations based on PC 1 is only possible when the MALDI-TOF MS score values (Figure 5.5 (D)) are used. The score values of FTIR (Figure 5.5 (A)) and Raman data (Figure 5.5 (B)) for example show very similar distributions for the two populations Sweden and Italy. In order to include more than one principal component when evaluating separation of the three populations by PCA, d-values were determined by MANOVA of the scores of the first ten principal components of each PCA/data block. For all data blocks, a d-value of 2 is obtained. This corresponds to the separation between three groups, here to three populations. Therefore, it can be concluded that a separation of the three different populations is possible with any of the data sets.

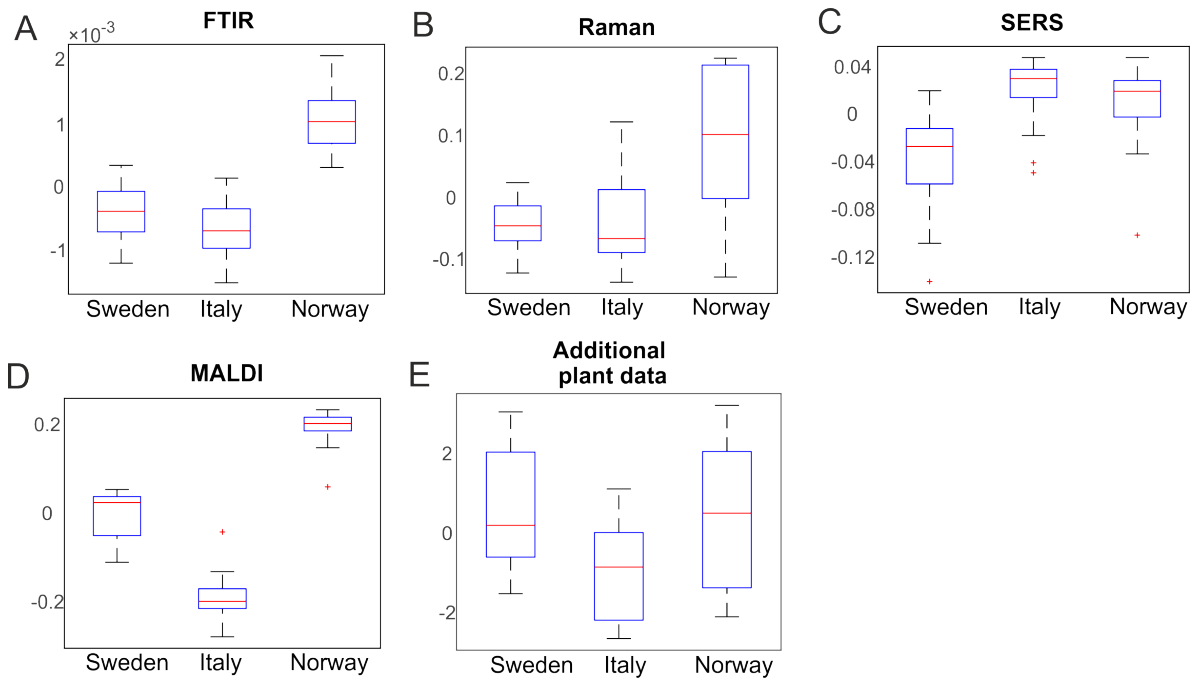

Figure 5.5: Results of the principal component analysis of **(A)** FTIR, **(B)** Raman, **(C)** SERS, **(D)** MALDI-TOF MS, and **(E)** additional plant data, respectively. Box plots display the variation of the score values of the first PC regarding each population obtained by Kruskal-Wallis H-Test.

The parent plants in each population were grown under four different conditions. Discrimination regarding potential effects of additional nutrients and temperature as design factors on pollen chemistry was studied using each of the five data sets separately as well. This was done for each population individually, as well as for all populations together. Table 5.1 summarizes the PCA results for each data block. The p-values were determined using PC 1 (Table 5.1, left column). In case of a high p-value when using the first PC, the lowest p-value with any of the other first ten PCs is shown in the table. The d-values were determined using the first ten principal components (Table 5.1, right column).

The first section of Table 5.1 displays the outcome of the PCAs obtained from the FTIR data sets. The separation based on FTIR spectra receives a p-value below 0.05 and a d-value of 3 for the populations Sweden and Italy, indicating that FTIR data enable differentiation of the applied growth conditions for these two populations. The FTIR data set of the population Norway with p-value larger than 0.05 and a d-value of 2 comprises less variance between growth conditions. When all populations are analyzed together, a high p-value for the first PC is obtained, which means that none of the four different growth conditions is separated using the variance explained by the first PC. Nevertheless, using the first to tenth PC, the d-value of 3 indicates a possible separation of all four growth conditions by FTIR alone.

Using the Raman data sets, the p and d-values of the PCAs from the data of the populations Sweden and Norway indicate a less sufficient discrimination ability (Table 5.1, second section). Only for the population Italy, a low p-value and a d-value of 3 can be interpreted as

a separation of the four groups from different growth conditions. In addition, the analysis of all populations together leads to a small p-value, showing the separation of at least one group based on Raman spectral information. The smaller discrimination ability compared to FTIR could be caused by the different selectivity of Raman spectroscopy. The high variances according to the growth conditions in the population Italy explained by the first PC are remarkable and in good agreement with studies by Zimmermann *et al.* on phenotypic plasticity in pollen.[15] The higher the phenotypic plasticity, the more the chemical composition in pollen varies when environmental conditions change. The high phenotypic plasticity of the population Italy has been inferred from FTIR spectra of the same *Poa alpina* population previously,[15] where a lower inner-group variance regarding different genotypes of the plants was found.

Investigation of the SERS spectra from aqueous pollen extract by PCA results in p-values above 0.05 for each individual population as well as the whole data set (Table 5.1, third section), clearly shows that an analysis of the samples by SERS alone will not be sufficient for the discrimination of pollen from parent plants that were grown under different environmental conditions. Nevertheless, according to the p-values found in PC 2 in population Sweden and PC 4 in population Italy (p-values in parentheses in Table 5.1), the variances from the effect of the growth conditions can also be detected in the aqueous extract for these two data sets and therefore add complementary information in the multi-block analysis discussed below.

Table 5.1: Results of the PCA (p- and d-values) for the discrimination of pollen samples from all populations and from the individual populations grown under different environmental conditions. The p-values are obtained for the score values of PC 1. In case of p-values above 0.05 in PC 1, the lowest p-value with any of the other first ten PCs and respective PC are shown in parentheses. For the calculation of d-values, the score values of the first ten PCs were used.

Method	Population	p-values for the separation of the pollen samples based on environmental conditions	d-values for grouping based on environmental conditions (max. 3)
FTIR	Sweden	<0.01	3
	Italy	0.035	3
	Norway	0.072 (0.011, PC 4)	2
	all	0.64 (<0.01, PC 6)	3
Raman	Sweden	0.51 (<0.01, PC4)	1
	Italy	<0.01	3
	Norway	0.36 (<0.01, PC 3)	2
	all	<0.01	2
SERS	Sweden	0.73 (<0.01, PC 2)	2
	Italy	0.37 (<0.01, PC 4)	2
	Norway	0.64[a]	0
	all	0.78 (0.014, PC 7)	0
MALDI	Sweden	0.62 (0.046, PC 3)	2
	Italy	0.012	1
	Norway	0.018	1
	all	0.98 (<0.01, PC 5)	1
Additional plant data	Sweden	<0.01	1
	Italy	<0.01	2
	Norway	<0.01	2
	all	<0.01	2

a no p-value below 0.05 can be found for the first ten PCs.

In MALDI TOF MS, data from population Sweden have a p-value above 0.05, whereas the p-values for the other two populations stay slightly below 0.05 (Table 5.1, fourth section). In contrast, based on the outcome for the whole population, it cannot be concluded that the chemical composition varies as a result of different growth conditions based only on one PC. The d-value of 1 (obtained using the first ten PCs), found for the whole data set as well as for population Italy and population Norway, can be interpreted as the formation of two distinct groups of MALDI spectra. This is in agreement with results discussed in Chapter 4, where the high discrimination ability between pollen from plants growing with and without additional nutrients using MALDI data from the same samples of *Poa alpina* could also be obtained. Since the discrimination takes place in the range m/z 5000-9000, we infer that the detected

signals belong to proteins and their derivatives from pollen nutrient storage.
The last section in Table 5.1 contains the p- and d-values for the analysis of additional plant phenotype data, namely height and number of flowering shoots, plant dry mass, and chlorophyll content. The p-values for each population and for the data set with all three populations combined are below 0.05, and we conclude that the variances regarding the separation of at least one specific group of scores from the other growth conditions are high. The d-values for the analyses of the data sets, however, are 2 or smaller, indicating that discrimination regarding all four growth conditions is not obtained.

The variation contributions of the different design factors, such as population, nutrients, temperature and their interaction as well as the contributions from individual variation, were calculated by Dr.Valeria Tafintseva with an approach underlying ANOVA-PCA and ASCA.[207–210] Figure 5.6, A shows the contribution of all possible design factors, that is, each type of variation for the whole data set of 72 spectra for each method. The variation contribution of the populations (Figure 5.6 (A, cyan bars)) is very large in the four spectroscopic/ spectrometric data sets, larger than the variation contribution due to the different growth conditions (Figure 5.6 (A, blue, orange and yellow bars)). Interestingly, and in agreement with previous work,[15] contribution of variation of the individual samples (Figure 5.6 (A first column, purple bars)) is similar magnitude to that introduced by changes in growth condition of the parent plant, and in the data sets from SERS and MALDI (Figure 5.6 (A second and third column, purple bars)), this contribution by individual variation is even larger.
Considering the data gathered from the parent plants, the largest variation contribution is the effect of the nutrient addition (Figure 5.6 (A, rightmost bar, orange coloring)), obviously having more consequences for the constitution of the plant itself than for the chemical make-up of the pollen. In addition, differences between phenotypic features of the plants in the different population are of a similar magnitude as variation due to individual differences. (Figure 5.6 (A, rightmost bar, cyan coloring)). The contribution of the residual variation (Figure 5.6 (A, green bars)) is relatively high for all data sets. In some cases such as Raman and SERS (Figure 5.6 (A, second and third column, respectively)), the residual variation contributes the most. This must be due to the type of experiments, which are in these cases much more prone to spectrum-to-spectrum fluctuation Moreover, the Raman and SERS data sets were collected over a course of several weeks, whereas MALDI and FTIR were high-throughput measurements obtained in one-preparation procedures. So, the big residual variation in SERS and Raman can be explained by the experimental variations.
In Figure 5.6 (B), relative variation contributions of the growth conditions, namely temperature, nutrients and the interaction of both factors are presented. Variations by these factors are emphasized by omitting population variation, individual variation, and residual variation. To calculate these, a variation of each factor was normalized by the sum of the variations for the three factors of interest. While the variation contribution of both, the temperature and

nutrients is high for the three spectroscopic methods FTIR, Raman and SERS, for MALDI, the variation of the nutrient factor is higher than the variation contribution of the temperature. The contribution of the different design factors to the total variation were also analyzed for each population separately (Figure 5.6 (C), Figure 5.6)(D), and Figure 5.7). As an example, Figures Figure 5.6 (C) and Figure 5.6 (D) show the outcome of the analysis for the population Italy. For the population Sweden (Figure 5.7 A and B), the overall variation contribution of the individuals (Figure 5.7 A, purple), is higher compared to the other populations, and contribution of variation due to the growth conditions is rather small.
This type of analysis helps understanding the underlying variation in the data introduced by different design factors and by other unwanted factors. PCA analysis and other multivariate data analysis techniques, if successfully working on the data, ensure that the amount of relevant variation available in the data is enough to discriminate between groups. As an example, although the different growth conditions contribute to only 10 % of the variation in the FTIR data from all populations (Figure 5.6 (A, first column)) a good discrimination of growth conditions using the first ten PCs is observed, yielding a d-value of 3 (Table 5.1, first section). This shows that the methods are powerful enough to focus on the relevant information in the data and the residual variation is not systematic. Regarding the hierarchical nature of the variance, the results of the ASCA approach are in good agreement with the results obtained by PCA. In data sets that show large contributions by different sources of variation, separation in a PCA is not unequivocal (see Table 5.1).

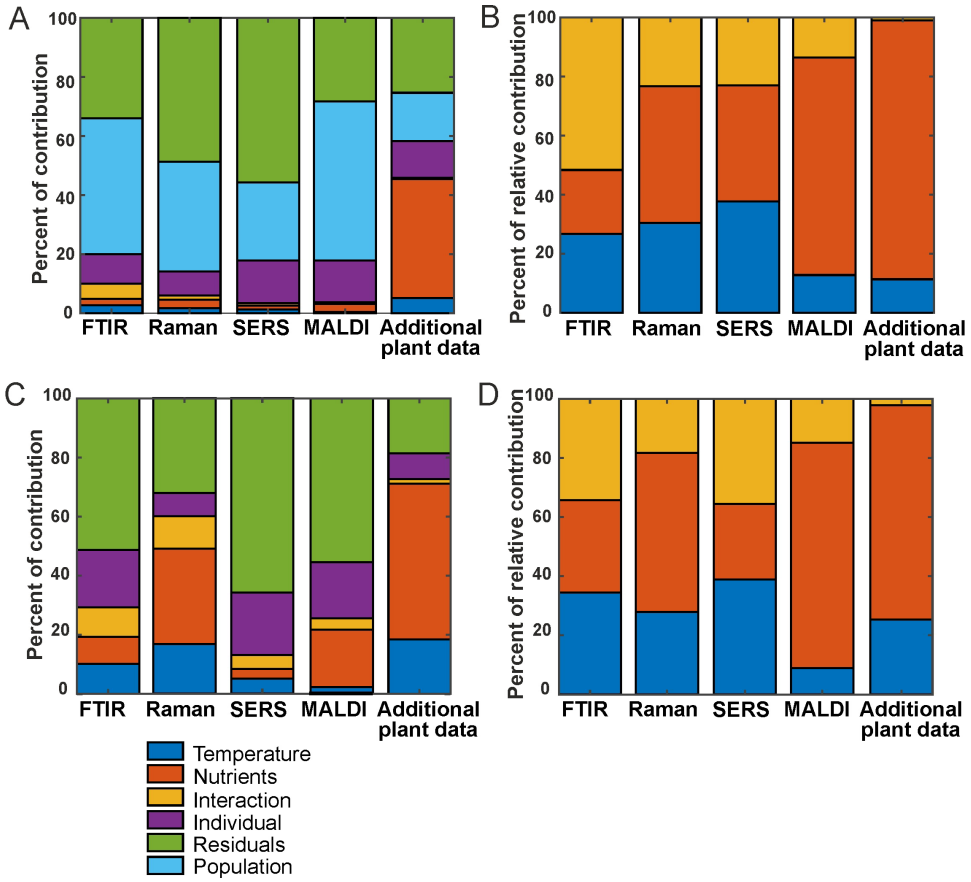

Figure 5.6: **(A)** Variation contribution of the design factors temperature (blue), nutrients (orange), the interaction of temperature and nutrients (yellow), different individuals (purple), populations (cyan), and residual variance (green) for the 72 spectra from the whole data set. **(B)** Relative contribution of temperature (blue), nutrients (orange), and the interaction of both (yellow) for the 72 spectra from the whole data set **(C)** and variation contribution of the design factors temperature (blue), nutrients (orange), the interaction of temperature and nutrients (yellow), individuals (purple) and the residuals (green) for the 24 spectra from the population Italy **(D)** Relative contribution of temperature (blue), nutrients (orange), and the interaction of both (yellow) for the 24 spectra from the population Italy. In (B) and (D), the contribution to the variance by specific population and the residual variance were left out, and the variation of each factor was normalized by the sum of the variations for the three other factors of interest. Variation contributions were calculated by Dr. Valeria Tafintseva.

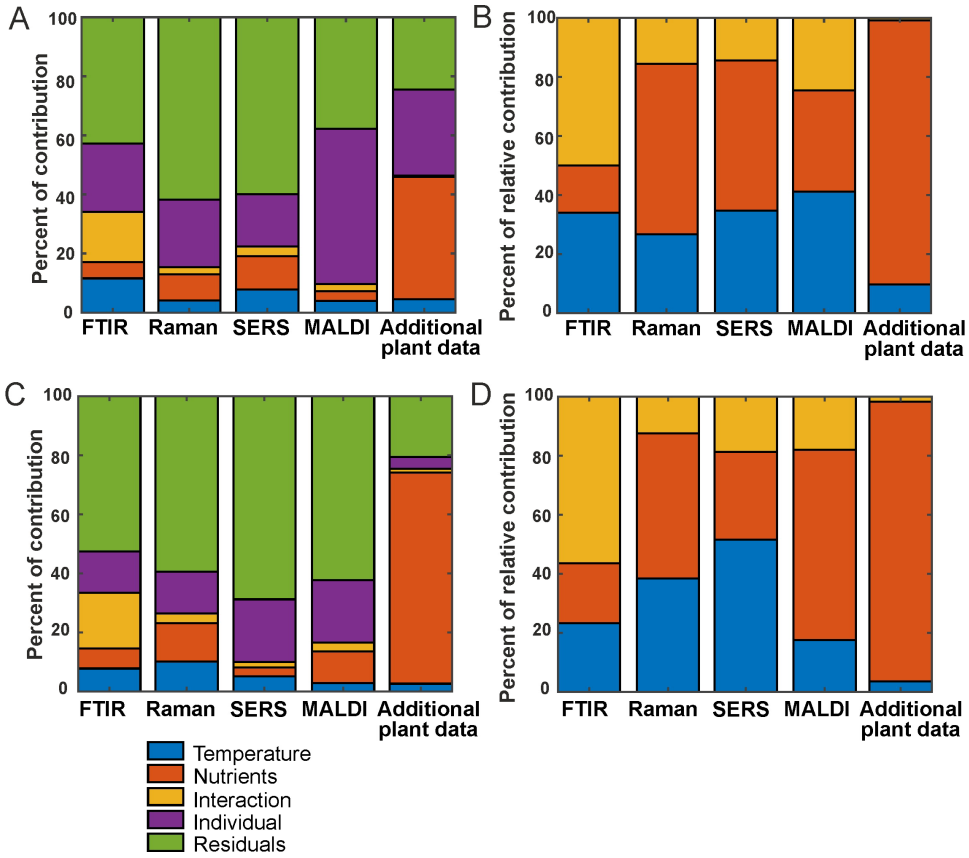

Figure 5.7: Variation contribution of the different design factors temperature (blue), nutrients (orange), the interaction of temperature and nutrients (yellow), individuals (purple) and the residuals (green) for the 24 spectra from the population Sweden (**A** and **B**) and for the 24 spectra from the population Norway (**C** and **D**). To emphasize the variations due to growth conditions, the individual variation and residual variation were omitted (B and D). Variation contributions were calculated by Dr. Valeria Tafintseva.

In conclusion, the different analytical methods vary greatly in their potential to discriminate the pollen from the sample set based on population and environmental influences. Due to the different selectivity in MALDI compared to FTIR, there is a superposition by the variation between the different genotypes (that is, individual variation) that impairs the discrimination ability for different growth conditions within one population. While both Raman microspectroscopy of single pollen grains and SERS enable classification of the pollen samples with respect to the corresponding population, no strong variation is found when these data sets are used to assess separation according to the varied environmental conditions of the parental plants. Nevertheless, the variation due to varied growth conditions is highly dependent on the considered population.

5.2 CPCA for the classification of pollen samples according to plant populations

With CPCA the five individual data blocks can be combined, and the impact of each method on the global analysis can be evaluated. Figure 5.8 shows the results of the CPCA for the classification of the different populations of *Poa alpina*, consisting of five block score plots (Figure 5.8 (B) to Figure 5.8 (F)) that correspond to the different analyses, and of a global scores plot (Figure 5.8(A)).
The global score values of the first and second CPC (Figure 5.8 (A)) show a clear discrimination of the three different populations. In particular, based on the variance represented by CPC 1, data from the population Norway and data from the population Italy are separated. As revealed by the block scores plots, the first component is mostly influenced by the FTIR block, comprising 41.7 % explained variance (Figure 5.8 (B)) and the MALDI block, explaining 39.62 % of the variance (Figure 5.8 (E)). The second PC is influenced in particular by the SERS data, explaining 37.55 % of the variance (Figure 5.8 (D)) and the block with the data on the parent plants, explaining 21.84 % of the variance (Figure 5.8 (F)). In all of the scores plots, the data sets of the population Sweden have positive score values, while the data sets of the populations Italy and Norway have mostly negative values regarding CPC 2 (Figure 5.8), particularly for the Raman (Figure 5.8 (C)), SERS (Figure 5.8 (D9)) and MALDI (Figure 5.8 (E)) block. The CPCA containing FTIR, Raman, SERS, and MALDI without the additional plant information leads to very similar scores plots, where also all three populations would be discriminated (Figure 5.9).

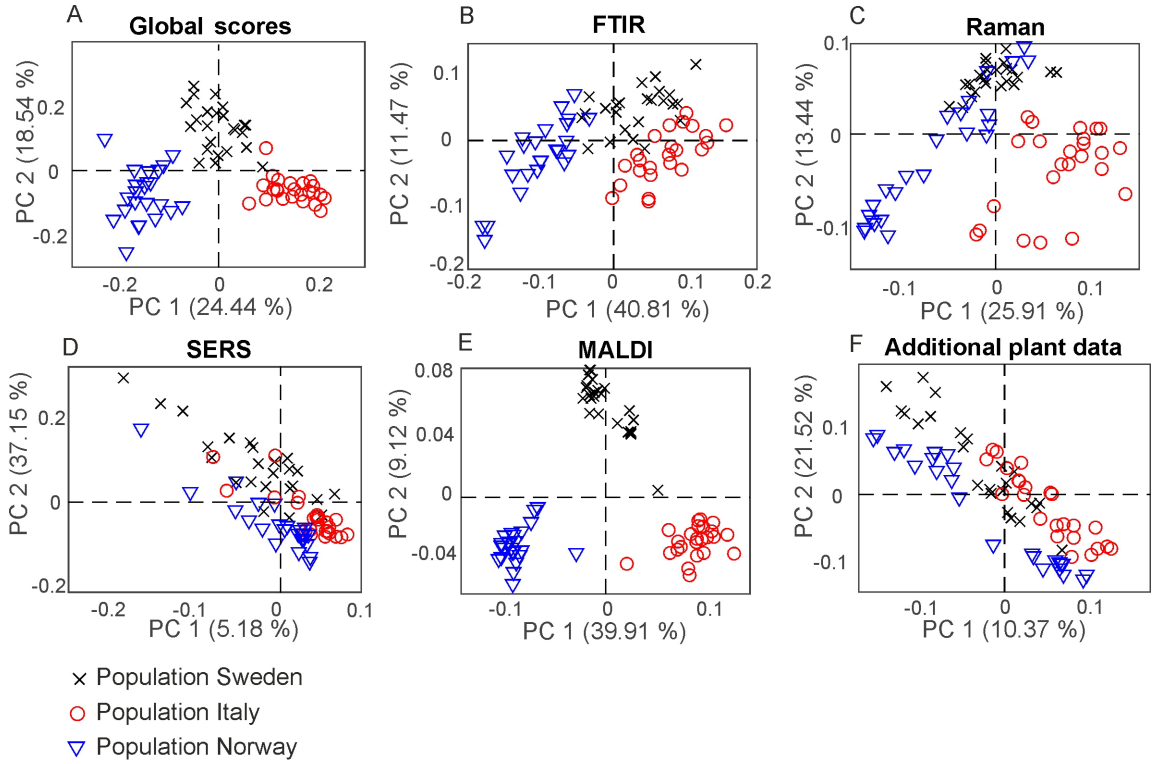

Figure 5.8: Score values of the CPCA analysis for the classification of samples from the populations Sweden (black crosses), Italy (red circles), and Norway (blue triangles). **(A)** Score plots for the global scores **(B-F)** individual data blocks. **(B)** FTIR, **(C)** Raman, **(D)** SERS, **(E)** MALDI, and **(F)** additional plant information.

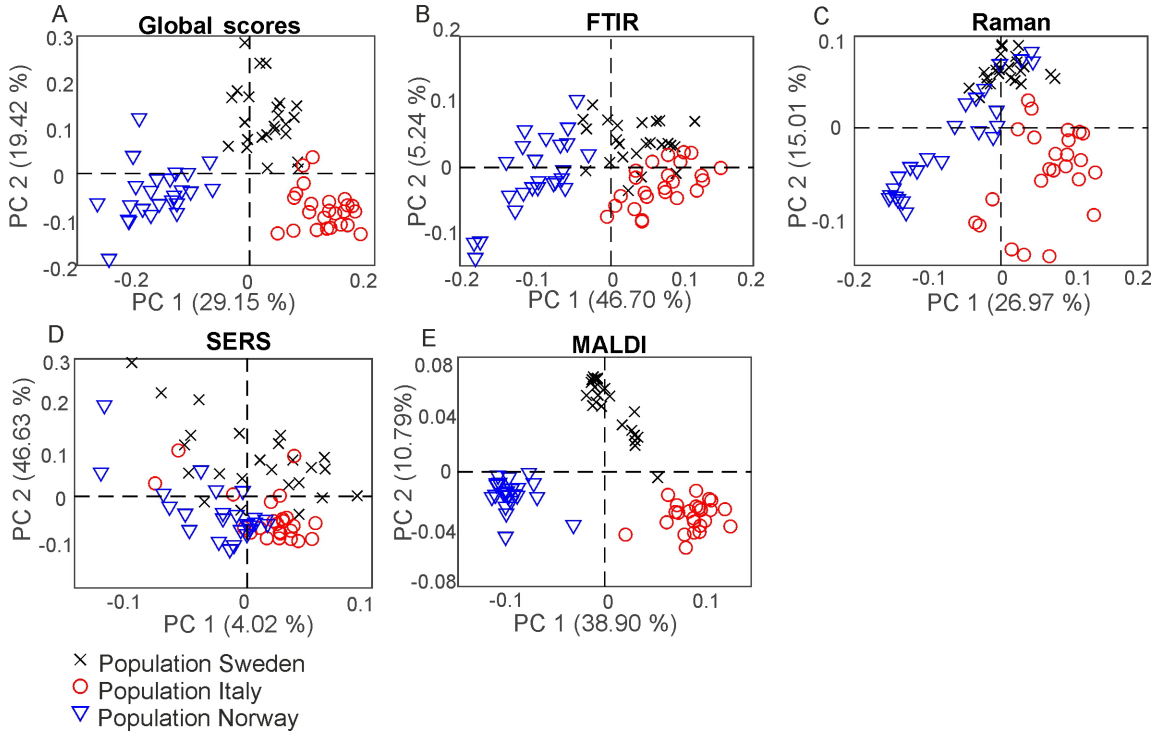

Figure 5.9: Scores of the CPCA analysis using the data blocks of FTIR, Raman, SERS, and MALDI without the additional plant information for the classification of samples from the grass pollen populations from Sweden (black crosses), Italy (red circles), and Norway (blue triangles). **(A)** Score plots for the global scores **(B-E)** individual data blocks. **(B)** FTIR, **(C)** Raman, **(D)** SERS, and **(E)** MALDI.

In Figure 5.10, the results of the separation of the respective first CPC are summarized in box plots for each block as well as for the global scores (Figure 5.10). Furthermore, a d-value of 2 was calculated based on the CPCA scores of CPC 1 to CPC 10 for the global scores as well as for all block scores. The data indicate that separation of the three populations is readily achieved based on the global scores (Figure 5.10 (A)), and that the FTIR (Figure 5.10 (B)) and the MALDI data sets (Figure 5.5 (E)) have the greatest influence on the separation in the global scores.

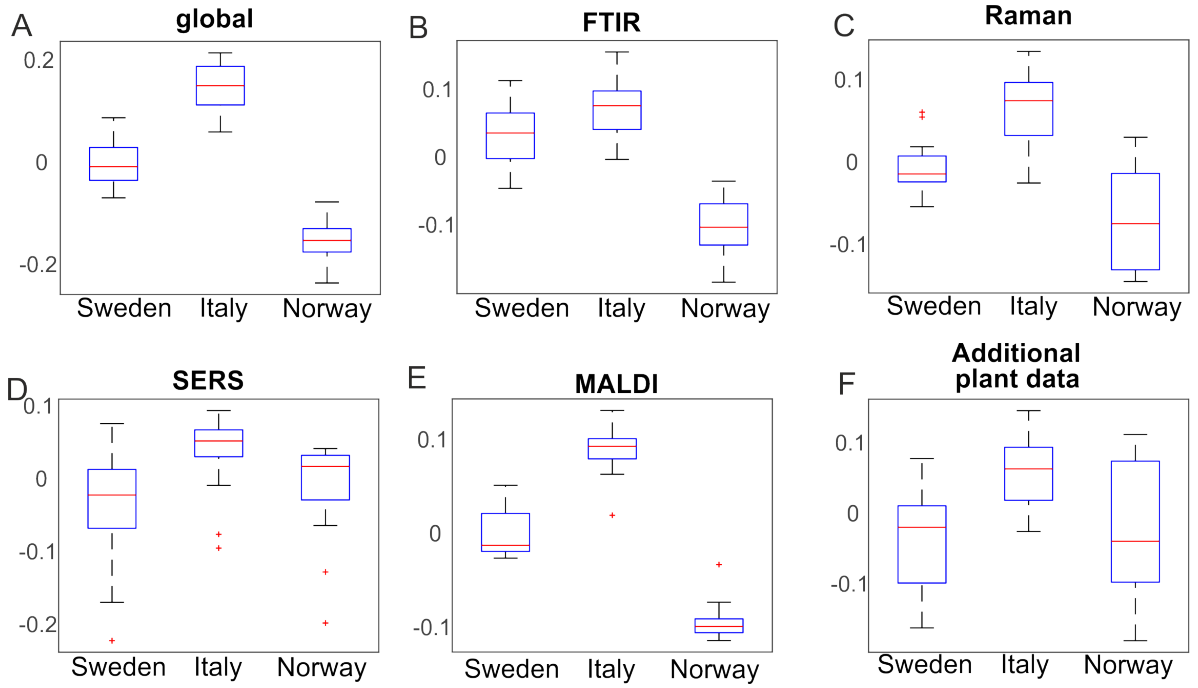

Figure 5.10: Results of the consensus principal component analysis of the five data sets visualized by box plots of **(A)** the global scores, **(B)** FTIR, **(C)** Raman, **(D)** SERS, **(E)** MALDI, and **(F)** additional plant information. The box plots display the variation of the score values of the first CPC regarding each population obtained by Kruskal-Wallis H-Test. Red lines indicate the median of the respective distribution, blue boxes represent the interquartile range, the black lines demarcate minimum and maximum values, and extreme values are shown as red markers.

In order to analyze which variables of the respective methods cause the separation in the global analysis and to investigate the correlations between them, a correlation loadings plot was generated (Figure 5.11). The plot shows the correlation between the global scores of the populations Sweden (red cross), Italy (red circle) and Norway (red triangle) and the relevant variables of the different blocks. For the clarity only the extrema of the loadings of the first and second component from the spectroscopic and MALDI blocks are shown, as well as all five variables from the additional plant data. Therefore, there are no variables visible close to the origin of the plot.

The different populations are characterized by variables that are located close to the global scores of the populations. The separation of the data from the population Sweden is caused by a high amount of spikes in the respective progenitor plants and their high dry mass. In addition, this population is characterized by Raman bands at 1007, 1161, and 1529 cm^{-1} that can be assigned to carotenoids[31] as well as by bands at 555 cm^{-1} that can be assigned to proteins,[13] and a MALDI peak at m/z 6038. The great influence that the SERS data block has on CPC 2, separating population Sweden (see Figure 5.8C), reflects in a correlation with SERS signals at 416, 733, 994, 1154, and 1545 cm^{-1} that are particularly important to discriminate the pollen data from the population Sweden (Figure 5.11, magenta markers). In the two other

populations, SERS signals at 581, 774, 1051, 1379, and 1424 cm^{-1} are observed. They might be attributed to the water-soluble part of proteins or carbohydrates.
The differentiation between the populations from Norway and Italy is achieved by utilizing CPC 1. The population from Italy is mainly separated by chemical information contained in the FTIR bands (Figure 5.11, blue markers) at 1026, 1079, 1151, 1472, 1525, 1649, and 1688 cm^{-1}, Raman bands (Figure 5.11, green markers) at 484, 649, 948, and 1609 cm^{-1}, and MALDI TOF MS peaks (Figure 5.11, yellow markers) at *m/z* 5282, 5968, 5980, and 6264. The FTIR and Raman bands can be assigned to starch, proteins, and sporopollenin vibrations.[5,10,13] Although an assignment of the MALDI peaks is more challenging, their positive correlation with these bands suggests that some of them are connected to nutrients, in agreement with previous discussions suggesting their assignment to oligosaccharides.[1,2]
The data sets of the population Norway show a positive correlation to the FTIR vibrational bands at 1089, 1166, 1503, 1666, and 1746 cm^{-1} as well as to the MALDI peaks at *m/z* 5880 and 6296 (Figure 5.11, bottom left section). Variances in Raman bands at 829 and 1043 cm^{-1} are positively correlated to the population Norway. Most of the Raman bands can be assigned to protein vibrations,[210] whereas the FTIR bands could be mainly assigned to carbohydrates.[10,15]
As illustrated by the band assignments, in addition to a redundancy in information (e.g., in some bands in FTIR and Raman spectra) each data block contains some exclusive molecular information, leading to their complementary. The different contribution of the five data blocks in the discrimination of the three populations shown in the correlation plot (Figure 5.11) indicates that particular parts of the pollen chemistry are responsible for the differences between populations, and that very different molecular/compositional parameters are responsible in the biochemical variation between two populations. The MALDI-TOF MS data have great influence on the analysis and can be exploited for a precise discrimination of all three populations. This is in accordance with the results of the PCA of the separate data block above (Figure 5.5 (D)) that indicates that MALDI-TOF MS and the biochemical fingerprint of glycoproteins and other macromolecules are specific for the pollen of a particular grass population.

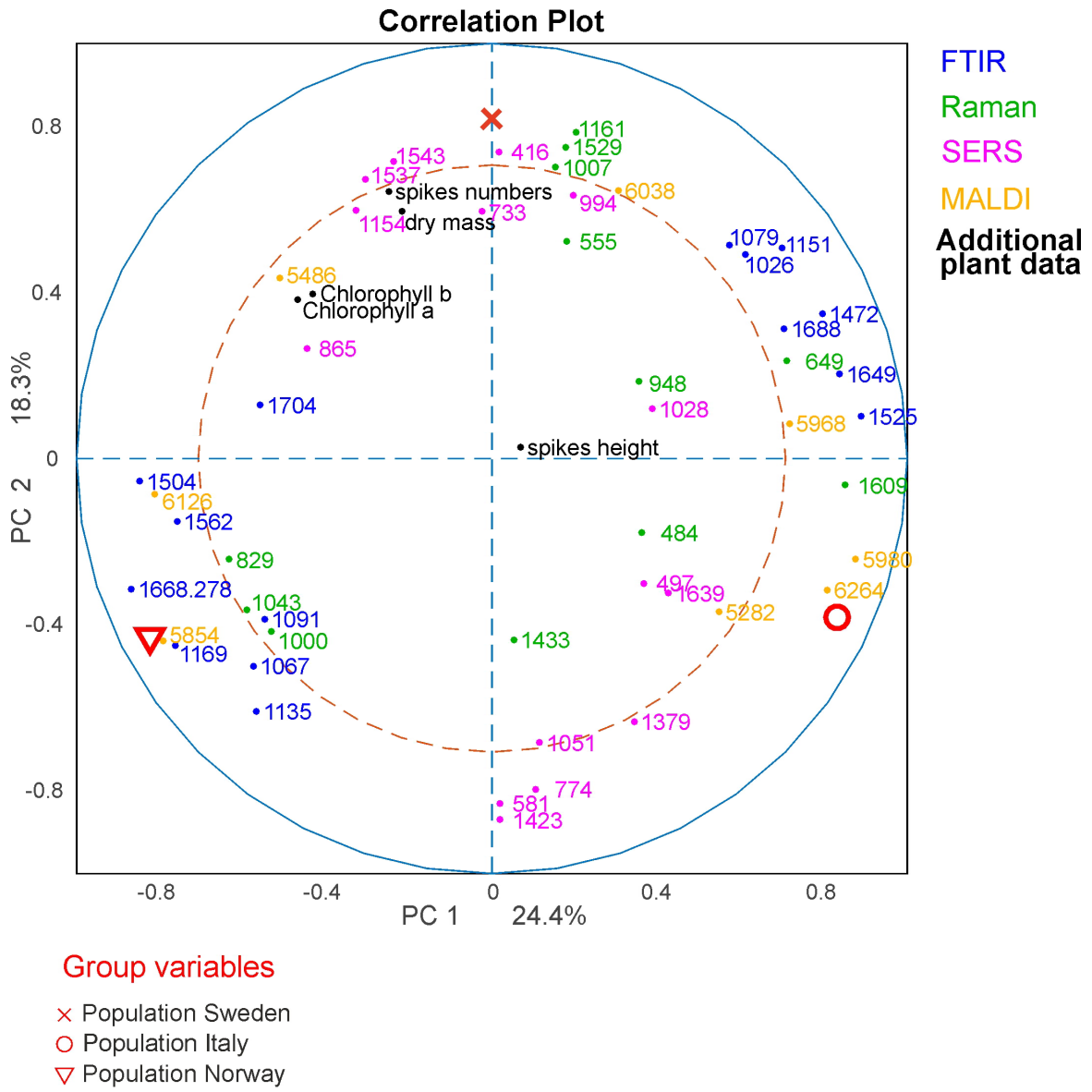

Figure 5.11: CPCA correlation loadings plot for the first and second CPC. The global scores of the three populations Sweden, (red cross), Italy (red circle), and Norway (red triangle), as well as the loadings of the blocks of FTIR, Raman SERS, MALDI-TOF, and additional plant data are displayed. For clarity only extrema of the loadings were shown for the spectroscopic/ spectrometric data.

5.3 CPCA for the classification of pollen samples according to different environmental influences

CPCA was applied to discriminate between pollen samples within each population that were collected from progenitor plants grown under four different environmental conditions: 14 °C and additional nutrients, 14 °C without additional nutrients, 20 °C and additional nutrients, 20 °C without additional nutrients. Table 5.2 shows the resulting p and d-values

analyzing the whole data set from all populations and the data from each of the three different populations individually for the global scores (Table 5.2, first section) and all the block scores (second to sixth section, respectively). The p-values for the global scores are below 0.05 for each population, indicating the separation of the different groups in the first CPCA component (Table 5.2, first section). However, considering all three populations together, separation is based on the third CPCA component. MANOVA utilizing the first ten CPCA components shows the highest possible d-value of 3, proving successful classification of all four groups of samples for population Italy, as well as for the whole sample set of all three populations. The lower d-value for the global scores in the population Sweden and Norway may be explained by a lower phenotypic plasticity of these populations compared to the population Italy.
Comparison of the results for the block scores (Table 5.2, second to sixth section) will help to identify those data blocks that are responsible for a separation based on the global scores. Based on the d-values, a separation of the samples into four groups - corresponding to four environmental conditions- is observed when all populations are analyzed together (last line in each of the sections of Table 5.2). The separation into four groups is possible for each of the five block scores except those of the MALDI block (last line in section 5 of Table 5.2). The Raman block scores indicate separation of the four groups in the two populations Norway and Italy (Table 5.2, third section). For the other block scores (FTIR, SERS, and MALDI), the separate analysis of each of the populations gives very different results. The samples from the population Italy showing separation according to the four growth conditions (Figure 5.1) in most of them, but less than four distinct groups in the populations Sweden and Norway. The block scores for the data gathered from the parent plants show very similar behavior and result in clear classification of all four conditions only in population Italy (Table 5.2, sixth section).

Table 5.2: Results of the CPCA (p- and d-values) for the discrimination of pollen samples from all populations and from the individual populations grown under different environmental conditions. The p-values are obtained for the score values of PC 1. In case of p-values above 0.05 in PC 1, the lowest p-value with any of the other first ten PCs and respective PC are shown in parentheses. For the calculation of d-values, the score values of the first ten PCs were used.

Method	Population	p-values for the separation of the pollen samples based on environmental conditions	d-values for grouping based on environmental conditions (max. 3)
Global	Sweden	0.036	2
	Italy	<0.01	3
	Norway	<0.01	2
	all	0.95 (<0.01, CPC3)	3
FTIR	Sweden	0.032	2
	Italy	0.013	3
	Norway	0.084 (0.021, CPC4)	3
	all	0.70 (0.011, C PC3)	3
Raman	Sweden	0.47 (0.013, CPC6)	2
	Italy	<0.01	3
	Norway	0.19 (<0.01, CPC5)	3
	all	0.64 (<0.01, CPC3)	3
SERS	Sweden	0.093 (0.032, CPC5)	2
	Italy	0.089 (<0.01, CPC3)	3
	Norway	0.40[a]	3
	all	0.98 (<0.01, CPC3)	3
MALDI	Sweden	0.60 (0.027, CPC5)	2
	Italy	<0.01	3
	Norway	<0.01	3
	all	0.98 (<0.01, CPC3)	2
Additional plant data	Sweden	<0.01	2
	Italy	<0.01	3
	Norway	<0.01	2
	all	<0.01	3

a no p-value below 0.05 can be found for the first ten PCs.

The weighting of each block in the CPCA can be interpreted and allows more insight into the influence of the data blocks on each other. As an example, Figure 5.12 shows the results for the CPCA applied to the data of the pollen samples from the population Italy. The first component of the global scores plot (Figure 5.12 (A)) separates between positive scores values of the data of samples from progenitor plants that were grown with the addition of nutrients (black crosses and blue triangles) and negative score values for samples from plants that were grown without the addition of nutrients (red circles and green diamonds). In Figure 5.12 (C)

and Figure 5.12 (F), the great influence of the Raman and the additional plant data block are revealed. Both blocks display similar group formation in the scores plots with high variances explained by the first CPC of 35.52 % and 54.50 %, respectively. The corresponding p-values in Table 5.2 are very low.

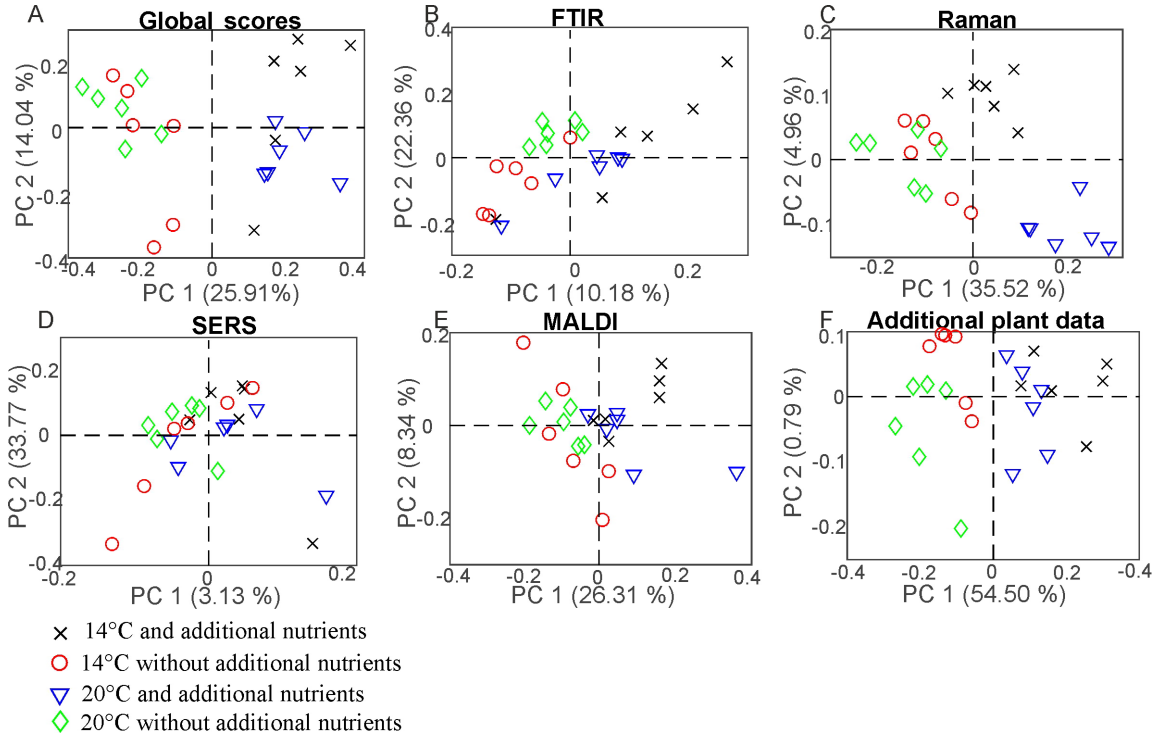

Figure 5.12: Scores of the CPCA analysis for the classification of samples from pollen of the population Italy regarding the four different growth conditions 14 °C and additional nutrients (black crosses), 14 °C without additional nutrients (red circles), 20 °C and additional nutrients (blue triangles), and 20 °C without additional nutrients (green diamonds). (**(A)** Score plots for the global scores **(B-F)** individual data blocks. **(B)** FTIR, **(C)** Raman, **(D)** SERS, **(E)** MALDI, and **(F)** additional plant information.

The scores of the second CPCA component separate pollen samples grown at 14 °C , as well as at 20 °C without additional nutrients (black crosses, red circles, and green diamonds) with positive values from negative values of those pollen samples grown at 20°C without additional nutrients (blue triangles) (Figure 5.12 (A)). The CPC 2 is mainly influenced by the SERS data (Figure 5.12 (D)), explaining 33.77 % of the variance. In the plot of the block score values (Figure 5.12 (D)), no separation of the groups that could correspond to growth conditions of the plants can be found. This suggests that other sources of variance, in this experiment resulting from individual genotypes, superimpose the influence of the growth conditions as discussed in Chapter 4. It is also in agreement with the calculated p- and d-values for the SERS block (Table 5.2, section 4). Furthermore, the Raman block scores plot, as well as the scores from the additional plant data, show great potential regarding the discrimination of different growth conditions in the population Italy. Since the additional plant data block explains most of the variance in the first CPC, CPCA was also performed without it, by using

only the spectroscopic/spectrometric data blocks, in order to confirm that the obtained global pattern is also driven only by the pollen chemical composition (compare with Figure 5.13), not by phenotypic features of the parent plant. Nevertheless, the additional plant data lead to a more complete view in this study and show correlation to the spectroscopic data blocks (compare Figure 5.14).

The molecular differences that cause the separation of the data show in the correlation loadings plot for the data from population Italy (Figure 5.14). Again, only those loadings with the highest impact are shown for clarity and the variables of the additional plant data were presented in full. As expected after the discussion of the block scores (Figure 5.12), the first CPCA component that separates samples from plants grown with additional nutrients (crosses and triangles) from samples without additional nutrients (circles and diamonds, also compare Figure 5.12) is mainly influenced by the Raman block and the additional plant data. Raman bands that characterize pollen samples with nutrient addition are 474, 830, 1003, 1435, and 1602 cm^{-1}. The bands at 1435 cm^{-1}, and 1602 cm^{-1} can be assigned to lipids[12,13] and a higher mitochondrial activity, respectively.[135,204] The other bands are associated with proteins.[10,13] The negative scores of the first CPC and the data of the pollen samples without additional nutrients (Figure 5.14, diamonds and circles) are mainly influenced by Raman bands at 485, 949, 1010, 1138, and 1471 cm^{-1}. These bands are associated with carbohydrates, such as starch.[13,123,203] Pollen are storing their nutrients in lipid bodies as well as in starch bodies, which are occupying most of the space in pollen grains.[25] The results confirm that plants growing under different nutrient conditions vary in their quality or amount of such storage bodies inside the pollen.

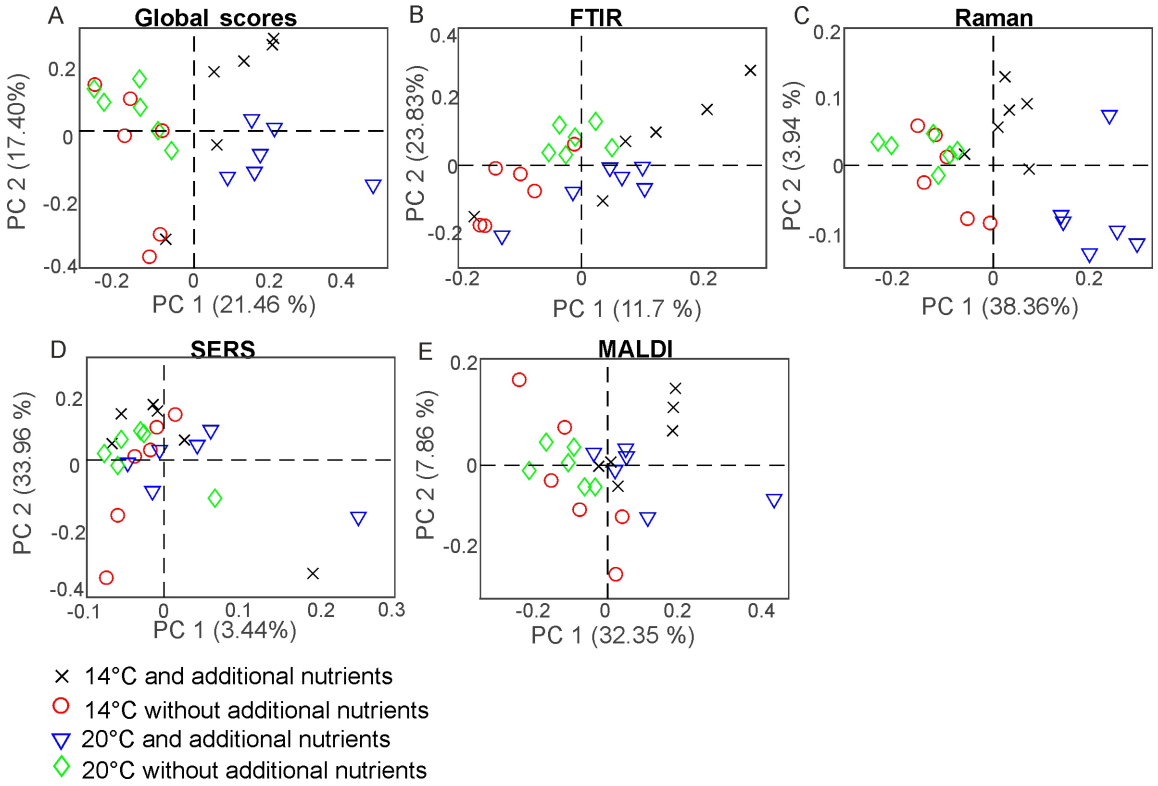

Figure 5.13: Scores of the CPCA analysis for the classification of samples from the grass pollen the population Italy regarding the four different growth conditions 14 °C and additional nutrients, (black crosses); 14 °C without additional nutrients, (red circles); 20 °C and additional nutrients (blue triangles); 20 °C without additional nutrients (green diamonds). **(A)** Score plots for the global scores **(B-F)** individual data blocks. **(B)** FTIR, **(C)** Raman, **(D)** SERS, and **(E)** MALDI.

CPC 2 can be used to separate between rather positive scores values corresponding to samples that were grown at low temperatures (crosses and circles) and rather negative scores values corresponding to samples that were grown at 20 °C (diamonds and triangles). As discussed before (see Figure 5.12) this separation is mainly influenced by SERS and FTIR bands. In particular, samples from plants grown at lower temperatures are characterized by a set of SERS bands including 445 cm^{-1} and the FTIR bands at 1721 and 1475 cm^{-1}. The FTIR bands can be assigned to lipids.[10] Samples grown under higher temperatures are characterized by SERS bands at 419, 929, 957 and 1564 cm^{-1}, and FTIR bands at 1666 and 1503 cm^{-1}. The bands could be assigned to nucleobases and proteins.[3] The combination of SERS and FTIR data, it can be assumed that the discrimination regarding the different growth condition is probably mostly influenced by the chemical composition of the pollen interior, although - in the preparations for SERS experiments - also water soluble compounds from the pollen outer shell may be found in the aqueous extract.

Variances in different nutrient conditions are mainly influenced by Raman bands that can be assigned to pollen outer shell and nutrient storage, as well as by plant parameters that are present in the additional plant data block. The differences in amount and quality of lipid and

starch bodies inside the pollen grains are most likely responsible for a distinction of samples from plants grown at different nutrient conditions. This is in good agreement with studies on *Poa alpina* using only FTIR spectroscopy by Zimmermann *et al.*[15] The temperature conditions at which parent plants are grown mainly affects the SERS and FTIR data blocks, and, probably, mainly the chemical composition of the interior of the pollen grains. It has to be pointed out that this conclusion is only made based on the data of the pollen from population Italy, were the samples are showing the highest phenotypic plasticity of the three investigated populations. Within the other populations, the correlation of the signals can differ greatly, indicating higher phenotypic rigidity, as discussed above.

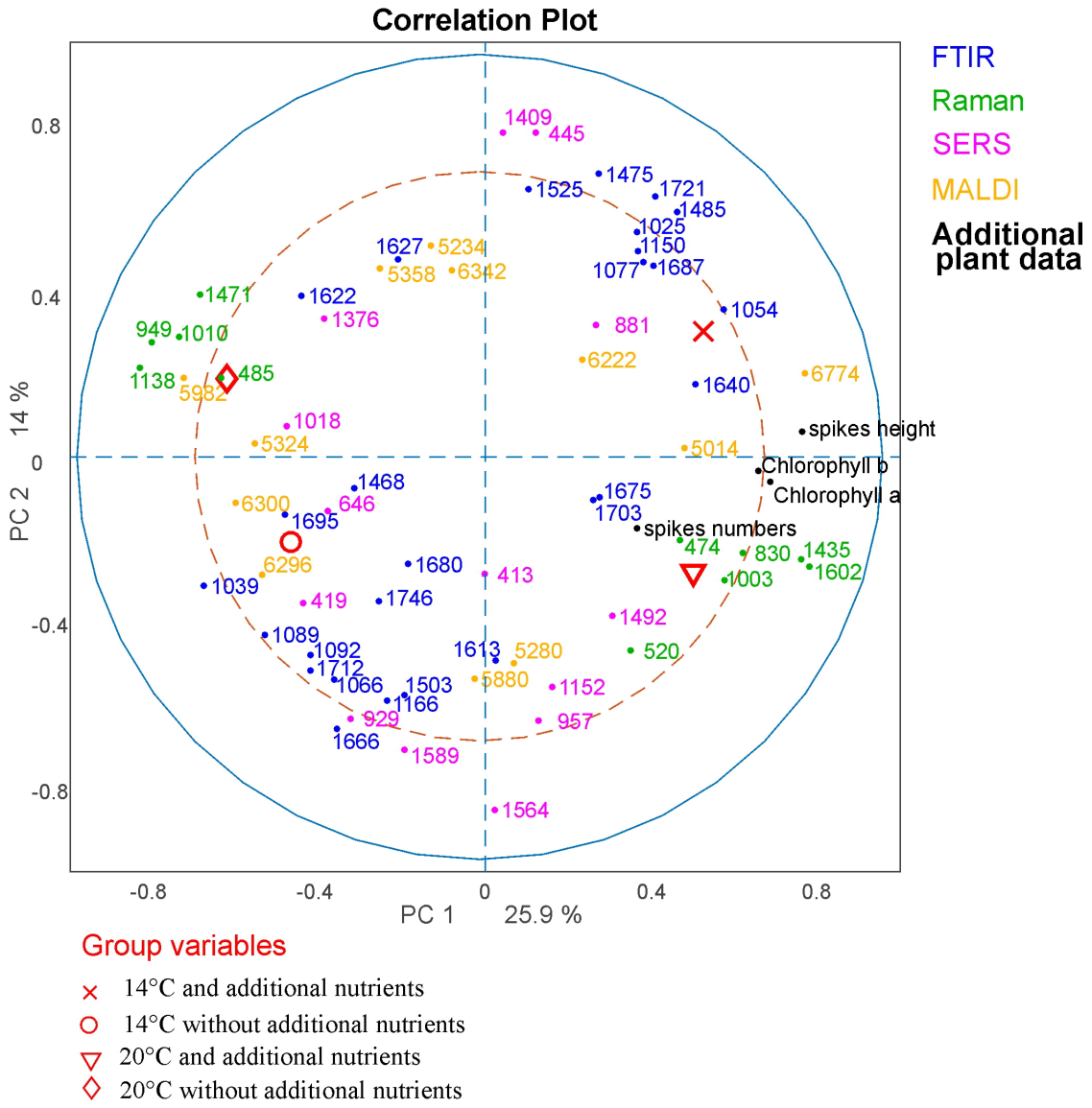

Figure 5.14: CPCA Correlation loadings plot for the first and second CPC. The global scores of population Italy regarding the four growth conditions 14 °C and additional nutrients, (black crosses); 14 °C without additional nutrients, (red circles); 20 °C and additional nutrients (blue triangles); 20 °C without additional nutrients (green diamonds), as well as the loadings of the blocks of FTIR, Raman SERS, MALDI-TOF, and additional plant data are displayed. For clarity only extrema of the loadings are shown for the spectroscopic/ spectrometric data.

In this chapter, four different spectroscopic and spectrometric methods, namely FTIR spectroscopy, Raman spectroscpoy, SERS, and MALDI TOF MS, are evaluated in their performance regarding the possibility to discriminate pollen with respect to their affiliated population and the growth conditions of their parent plants. Pollen samples of a subset from Pollen Norway I, containing 72 samples of *Poa alpina* provide variation in three populations and four growth conditions. Additional plant-related data serve as a fifth data block for multiblock analyis. The chosen methods are complementary regarding sample preparations, selectivity, and sensitivity of each analytical technique. Evaluation of the methods was conducted by PCA,

indicating that a separation of the three populations can be attained by each of the separate methods. Some of the methods are suitable to assess the variation due to different growth conditions. Particularly, the different populations can be discriminated easily by MALDI TOF MS and FTIR, whereas MALDI TOF MS is less efficient in the discrimination of different growth conditions.
A combination of the four spectrometric and spectroscopic methods,with additional plant data, such as the chlorophyll a and chlorophyll b content, amount of spikes, and spikes height that were used here, the possible correlation of biochemical pollen composition with plant morphology can be investigated. A similar problem, combining tissue biochemistry with other parameters in a plant is also discussed in Chapter 9. The population Italy, that shows high phenotypic plasticity, was investigated separately using CPCA. Using additional plant data, correlation between the biochemical pollen composition and plant morphology can be investigated.

6 Utilization of Raman spectra from single pollen grains

In this chapter, Raman mapping data of pollen grains from grass pollen will be used to investigate the following aspects:
(I) Based on the results discussed in Chapter 5, indicating the successful combination of different methods, substrates that allow the collection of Raman and MS data from pollen on the same substrate are desirable. Specifically, the advantages of sample preparation by a fixation on carbon tape and without fixation on calcium fluoride will be discussed. (II) Different from the Raman approaches discussed so far, here, also the feasibility to use Raman mapping data rather than point measurements will be investigated: Successful utilization of Raman data about the detection of variances within the same pollen species, e.g. for different populations and growth conditions, was demonstrated in the previous chapter using averaged spectra of each pollen sample. However, due to the high spatial resolution of Raman experiments compared to FTIR or Matrix assisted laser desorption/ionisation time of flight mass spectrometry (MALDI TOF MS), the heterogeneity of the pollen grain could lead to great fluctuations within the data set from one sample.[16,31] A larger total amount of spectra from different areas of the pollen grain can be collected by Raman microspectroscopy to reduce this variance.
(III) Variance in the chemical composition of pollen grains in one species does not necessarily have to be hierarchical as described in the experiment discussed in Chapter 5, but can be more complex. The discrimination between different pollen species and mutants as well as between different growth conditions of one species is investigated.
(IV) Study the influence of MALDI MS spectral information on the classification based on Raman mapping data.

6.1 Discrimination of Raman microspectra from different pollen species on calcium fluoride and carbon fixation tape

Pollen grains from the three different grass species *Anthoxanthum odoratum, Festuca ovina*, and *Poa alpina* were mapped using Raman spectroscopy. In multimodal measurements of pollen grains, a fixation would be suitable. It was presented, that fixation using carbon tape yield to discrimination of different pollen species in MALDI.[107] To compare the fixation methods also in Raman experiments, the three pollen species were measured on calcium fluoride and carbon tape using an excitation wavelength of 532 nm.

Figure 6.1 shows one of the pollen measurements using calcium fluoride and the 532 nm excitation. Two pollen grains from *Poa alpina* are mapped with 5 μm distance between the spots. In Figure 6.1(A and B), the microscopic image and the chemical image (1550 and 1700 cm^{-1}) are presented. Moreover, several background spectra were measured using a rectangular map geometry, yielding a strong contrast between the pollen spectra and background spectra in the chemical map. Figure 6.1 (C) shows one raw pollen spectrum (black) and one raw spectrum from the calcium fluoride(blue). The pollen spectrum shows a strong fluorescence background (Figure 6.1 (C), black). In the raw spectra of the pollen region, three bands can be seen at 1000, 1252, and 1575 cm^{-1} which can be assigned to carotenoids.[13] In contrast, the background spectrum (Figure 6.1, D) does not show any signals since the CaF_2 bands are not in the considered spectral range.

For the discrimination of pollen grains from different grass species, three pollen grains from each pollen species were measured and analyzed using PCA. The results are shown in Figure 6.2 (A). No separation of the different pollen species is observed. The highest variance is based on the carotenoid signals at 1000, 1160, and 1524 cm^{-1}, correspond to the discrimination between the background and the pollen spectra (Figure 6.2 A, loadings). Therefore, the background spectra need to be separated from the pollen spectra (Figure 6.2 B).

The PCA on the extracted pollen spectra shows no discrimination between the three different species in PC 1 and PC 2(Figure 6.2 B). The main variance is the variation within the carotenoid signals, which can decrease during the measurements due to the photo-bleaching effect.[13] Nevertheless, score values of the data from *Anthoxanthum odoratum* are less spread out and show mostly negative values for PC 1 and PC2. Two positive carotenoid signals at 1164 and 1534 cm^{-1} and also two negative signals at 1154 and 1515 cm^{-1}, that can be assigned to carotenoids as well. It could be interpreted, based on excitation with 532 nm, that the pollen differ slightly in their carotenoid composition.

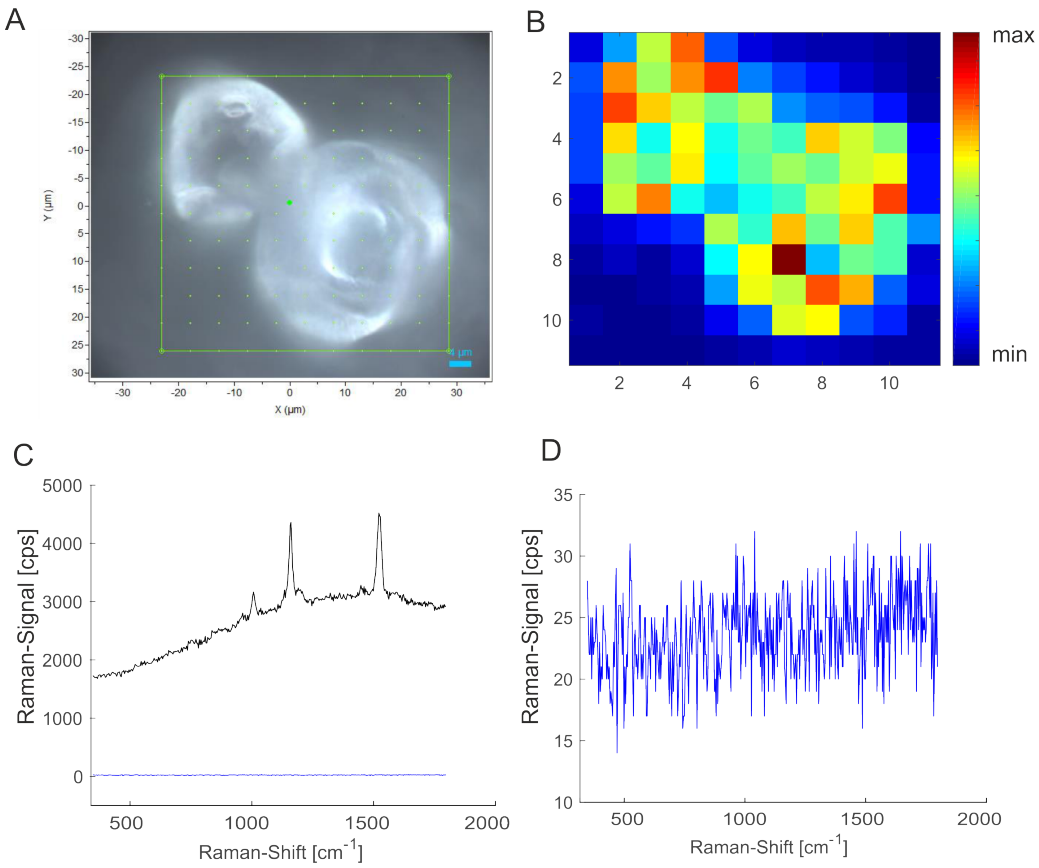

Figure 6.1: Measurements of pollen grains on calcium fluoride. **(A)** Microscopic image of the mapped pollen grains using a step-size of $5\mu m$. **(B)** Chemical image over the spectral range from 1550 to 1700 cm^{-1}. **(C)** Representative spectrum from the pollen grain, (black) and the calcium fluoride substrate (blue). **(D)** Magnification of the blue spectrum in C.

For comparison, Figure 6.3, A shows a pollen grain from *Festuca ovina* measured on carbon tape and the corresponding chemical image. The pollen spectra show the three carotenoid bands comparable to the spectrum in Figure 6.1 (C). In Figure 6.3 (C), one raw spectrum from the pollen grain (black) and one spectrum from the background are presented. The magnification of the background spectrum (Figure 6.3, D) indicates that two additional bands from carbon are visible at 1354 and 1595 cm^{-1}.[211] The two bands have a low intensity compared to the high fluorescence signal in the specific sample here, but could influence the vibrational signature, when fluorescence is low, e.g., after pre-processing or at different excitation wavelength.

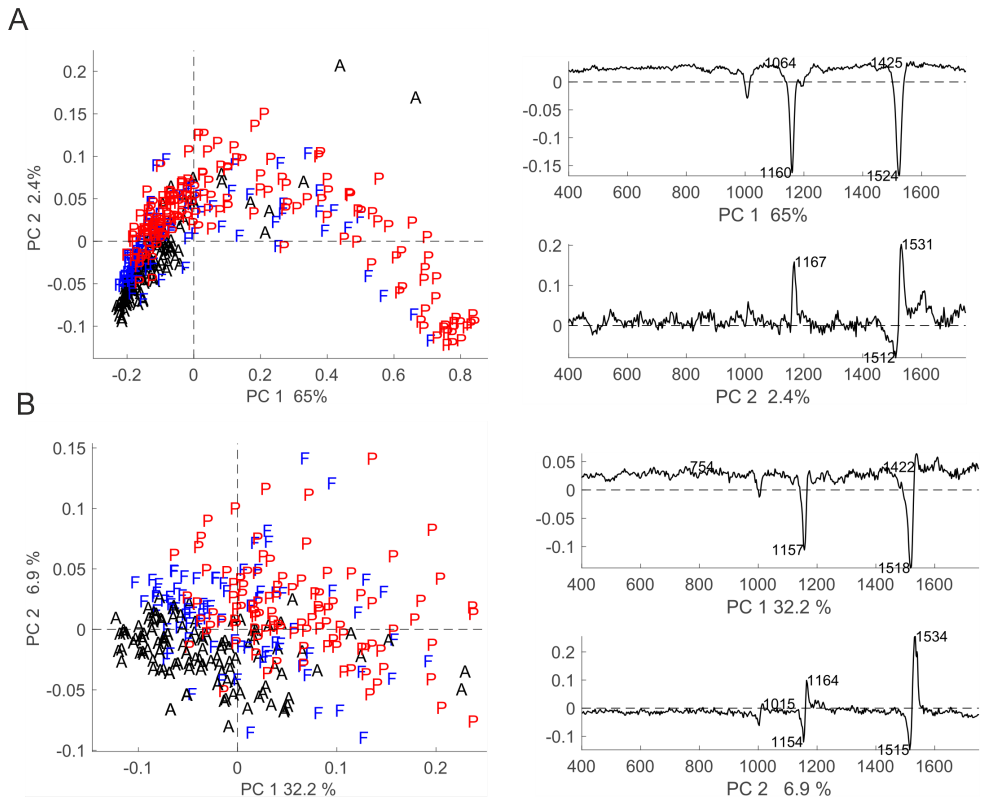

Figure 6.2: **(A)** Scores plot and corresponding loadings for the first and second PC of all 408 spectra and **(B)** 316 extracted spectra. Black, *Anthoxanthum odoratum*, blue, *Festuca ovina*, red, *Poa alpina*. Spectra are pre-processed using interpolation in the range of 400-1700 $cm-^{-1}$, AsLS correction and vector normalization. Extraction was executed using hierarchical cluster analysis as discussed in Chapter 9 in more detai.l

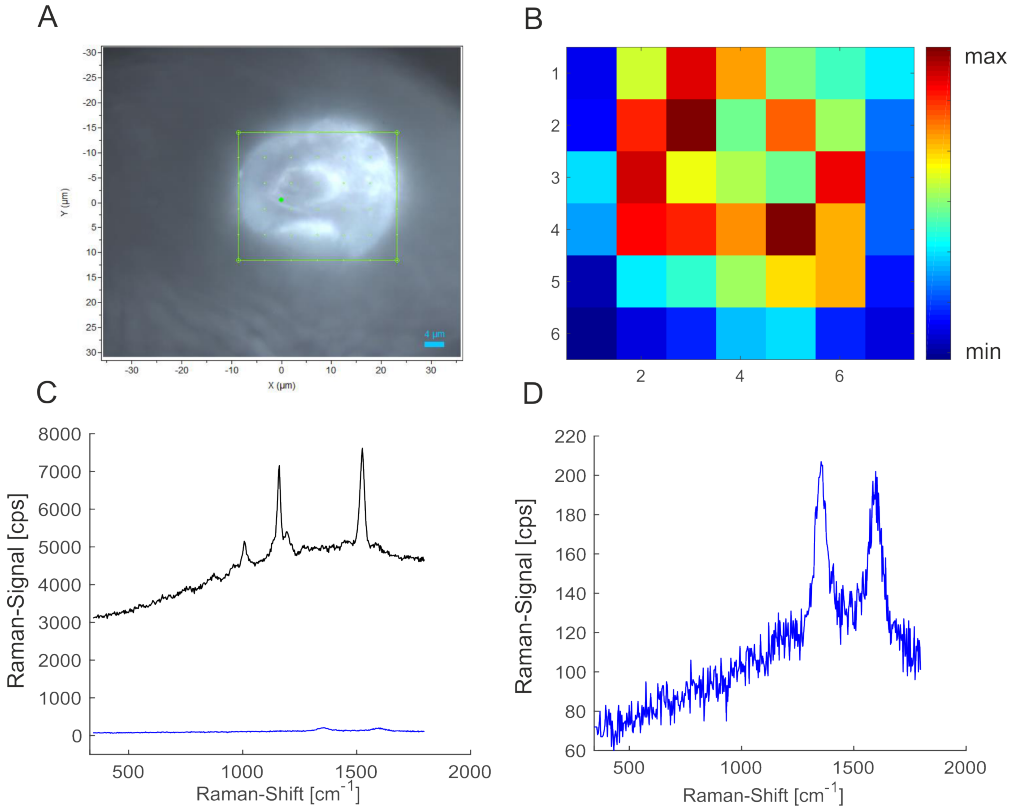

Figure 6.3: Measurements of pollen grains on carbon tape. **(A)** Microscopic image of the mapped pollen grains using a step-size of 5 μm. **(B)** Chemical image over the spectral range from 1550 to 1700 cm^{-1}. **(C)** Representative spectrum from the pollen grain, (black) and the calcium fluoride substrate (blue). **(D)** Magnification of the blue spectrum in C.

The two different preparation methods, measurement on calcium fluoride and fixation on carbon tape, are compared with each other in the PCA presented in Figure 6.4. Here, only pollen spectra from the same species *(Festuca ovina)* were taken into account. The scores plot indicates a separation according to the carbon tape and calcium fluoride approach, mainly based on the carotenoid composition. There are no qualitative differences in the corrected spectra apart from the carotenoid bands, as can be seen in Figure 6.4 (B).
In general, the measurement of pollen on carbon tape would be possible. Unfortunately, only the measurement using 532 nm excitation would lead to spectra, which limits the discrimination ability of several pollen species, since the fluorescence and carotenoid bands are dominating the spectra. Some pollen species have different carotenoid composition so that there is still the possibility to discriminate based on carotenoid bands themselves, which would enable a possible multiblock approach with other spectroscopic or spectrometric methods, particularly MALDI MS, where carbon tape has been reported as useful substrate[107]

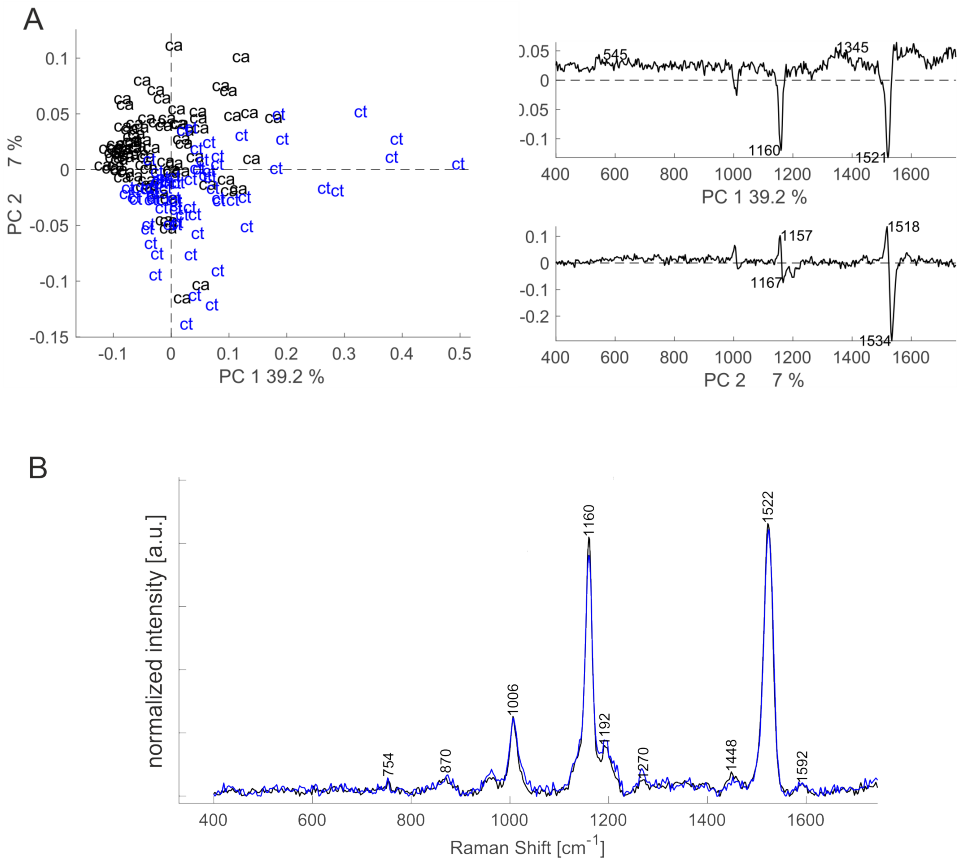

Figure 6.4: (A) PCA of 156 pollen spectra from *Festuca ovina* on calcium fluoride (Ca, black) and carbon tape (Ct, blue). **(B)** Comparison of a pre-processed spectrum from *Festuca ovina* on calcium fluoride (black) and carbon tape (blue).

6.2 Sampling and experimental conditions in Raman mapping experiment of pollen

Since in a Raman mapping experiment of pollen, a spatial heterogeneity of chemical composition is observed, changes or differences in pollen biochemistry that occur in collection and sampling may become visible more clearly than in the point measurements discussed so far (Chapter 5). For example, Figure 6.5 shows one pollen grain before and after a Raman map at 785 nm. Usually, the pollen measurements in Raman experiments should be invasive, but in some occasions, the pollen grains can also burning during measurements.

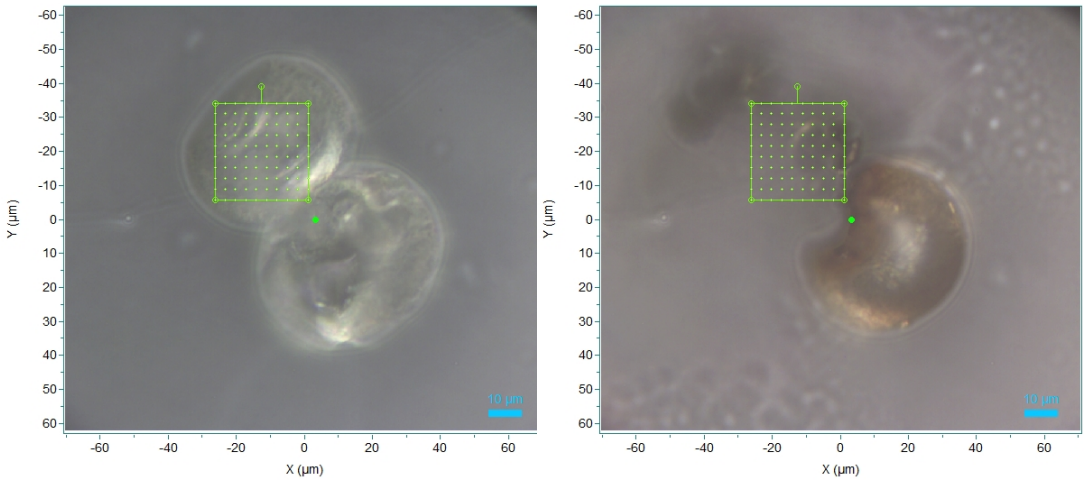

Figure 6.5: Bright-field image of a pollen grain from a *Sorghum bicolor* wild-type plant **(left)** before and **(right)** after a measurement.

As another example, the effects of unintended pollen germination will be discussed. It is well known that pollen ultra-structure, such as pollen tubes, can be analyzed using Raman spectroscopy.[13,16,153] Usually, the germination of grass pollen is difficult. A study on germination rates of the *Sorghum bicolor* pollen grains using light microscopy to differentiate between the wild-type and mutant as well as between control and stressed plants was executed as a bachelor thesis. As a result, only a few pollen tubes were found, but slight differences in the size distribution of pollen in medium and pollen in water are obtained (Niclas Schauer, Bachelor thesis, unpublished data). Pollen grains in water show less variation in size and shape than in the case of a medium containing sugar, boric acid, and potassium salts. Nevertheless, comparing the effect of the medium on the pollen grains no difference, neither in pollen from mutant or wild-type nor in the stressed and controlled plants.
In some cases, the pollen grains are probably germinated by chance, due to conditions like temperature changes while shipping the samples. No additional medium or incubation was applied in this case. As in the example shown in Figure 6.5, the likelihood of laser damage indicated by carbon formation in some of the samples was very high. This could possibly be explained by a different chemical composition of the pollen grains, known to occur during germination.[153,212] As a consequence, only a small part of the germinated pollen grain was measured to avoid the burning of the sample.
In Figure 6.6 the bright-field image of one measurement of a pollen grain and the pollen tube is shown as an example. The green area marks the mapping area.

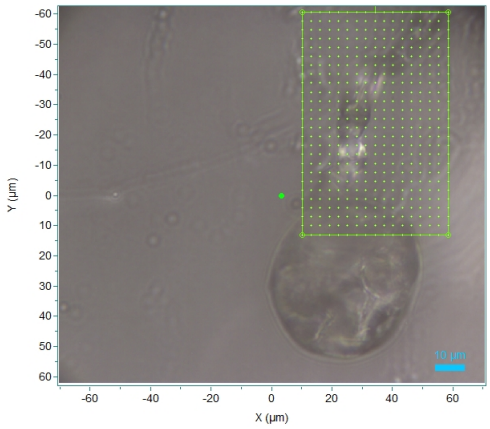

Figure 6.6: Bright-field image of a pollen grain with pollen tube. Scale bar, 10 μm.

Based on the intensities of different important bands, chemical images can be calculated from the Raman spectra. Here, the area between the spectrum and the baseline is calculated for the defined range (Figure 6.7). Figure 6.7 shows the five chemical images based on bands that can be assigned to specific chemical components of the pollen grain and pollen tube, namely starch (450-500 cm^{-1}),[10] proteins (1400-1500 cm^{-1}),[5] sporopollenin (1550-1700 cm^{-1}),[5,10] coniferyl aldehyde (1100-1300 cm^{-1}),[10] and on the two prominent bands of carotenoids (1100-1200 cm^{-1} and 1500-1600 cm^{-1}).[13] Figure 6.7 (A) shows the spatial distribution of carbohydrates as starch in the pollen tube. The band around 480 cm^{-1} is dominant in the spectra of pollen as it can be seen for example later in this section in Figure 6.12 or in literature.[13,16] However, the band is less pronounced in the pollen tube. The protein band at 1460 cm^{-1} (Figure 6.7 (B)) is also more intensive in the spectra of the pollen grain, but more dominant than the band at 480 cm^{-1}. The sporopollenin band (Figure 6.7 (C)) is less intensive in the pollen tube. The two carotenoid bands (Figure 6.7 (E)) are more dominant in the pollen grain region and show a similar distribution as the starch band (Figure 6.7 (A)).

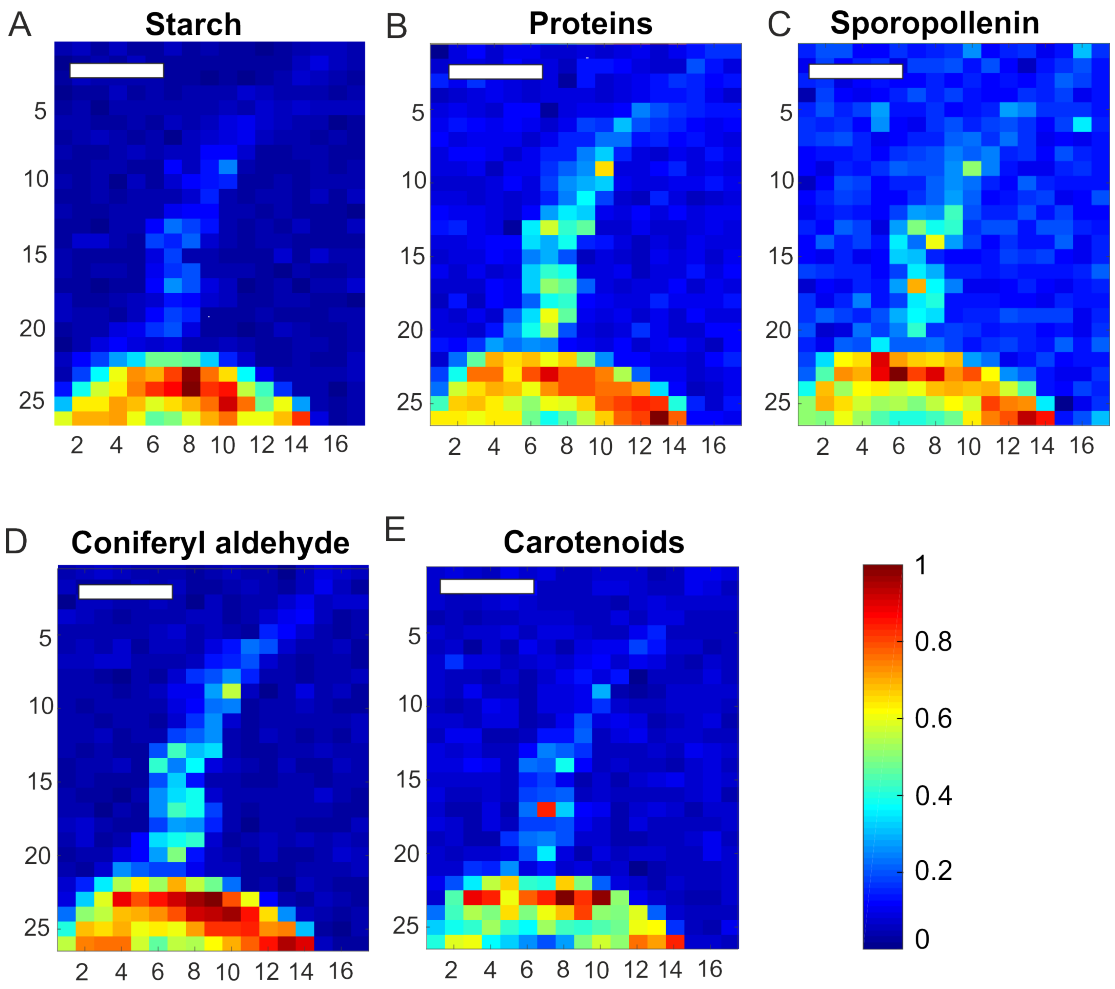

Figure 6.7: Chemical maps of the germinating pollen grain using selected bands (step size in x,y direction, μm,intensity, $1.4 \cdot 10^6$ $W cm^{-2}$ acquisition time per spectrum, 1 s, excitation wavelength, 785 nm). For mapping of the carotenoid component, two characteristic spectral ranges were used in order to enhance the probability to probe carotenoids. Spectra were interpolated in the respective area and baseline corrected with AsLS. The integrated intensity values were standardized for the respective min/max mapping. Scale bar,10 μm.

HCA is applied to form clusters of the spectra from the map. Figure 6.8 (left) shows the resulting HCA image with three clusters. One cluster (Figure 6.8, red) comprises the spectra of the pollen grain and the second cluster includes the spectra of the pollen tube (Figure 6.8, blue). The largest cluster includes spectra of the background that consists of a three-band pattern that is most likely caused by impurities of the calcium fluoride substrate.
In Figure 6.8 (right), the averaged spectra of each cluster are drawn in the specific color. Comparing spectra from cluster Pollen grain (Figure 6.8, red) and cluster Pollen tube (Figure 6.8, blue) several bands are missing in the pollen tube spectrum. Most obvious, the bands at 478, 942, and 1084 cm^{-1} are missing in the spectra from the pollen tubes (Figure 6.8, blue). Also, the band at 1342 cm^{-1} is more pronounced in the averaged spectrum of cluster Pollen grain (Figure 6.8, red). These bands can be assigned to carbohydrates, such as starch or pectin.[16,31] Thus, it can be concluded that the pollen grain has a higher starch content than the pollen

tube. This is not in agreement with previous investigations on pollen tubes by Schulte *et. al.*, and Joester *et al.*[16,153] In contrast to the unintended germination that occurred in the samples here in the absence of a germination medium, these studies[16,153] were carried out in optimized media and on pollen grains of a variety of species that, furthermore, excluded grass pollen. The tubes of grass pollen grains are known to contain carbohydrates, which play a role in the elongation.[212] This could also be part of an explanation of the chemical composition observed here. The unintended germination of pollen can play a role in many types of environmental samples, and should be considered in real applications of pollen analysis.

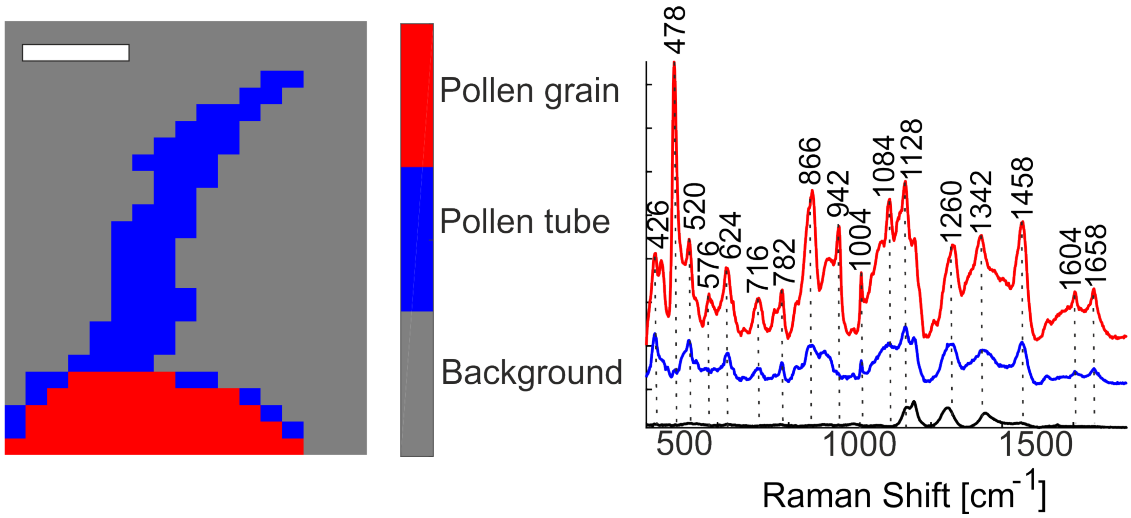

Figure 6.8: (Left) HCA image of the pollen tube using Euclidean distances and Ward's algorithm colored by the three biggest clusters and **(right)** corresponding averaged spectra.

6.2.1 Effects of spectral quality on the analysis of mapping data

The data set presented in this section includes 25 pollen samples from *Sorghum bicolor*. Figure 6.9 shows the framework of how the variances within the data are structured. The presented framework (Figure 6.9) deals with several biological questions: First, the discrimination between pollen from mutant plants and wild-type plants. The data set is divided into 11 wild-type samples and 14 mutant samples (Figure 6.9, green) that concern two different mutations, *SbLsi1* and (*bmr*. Mutation *SbLsi1* affects the silicon uptake[37] and the second one is related to a different kind of lignin (*bmr*[190]). This separation within the group of mutant plants would lead to a second branching in the framework (Figure 6.9, red). The wild-type plants as well as the *SbLsi1* - mutants are grown under different stress conditions,[191] resulting in seven control plants for wild-type and five control plants for mutants as well as four plants that grow up under stressed conditions (drought and salt stress). The third variation that is discussed here is the separation regarding pollen from plant which grow under normal and under stress condition(Figure 6.9, blue).

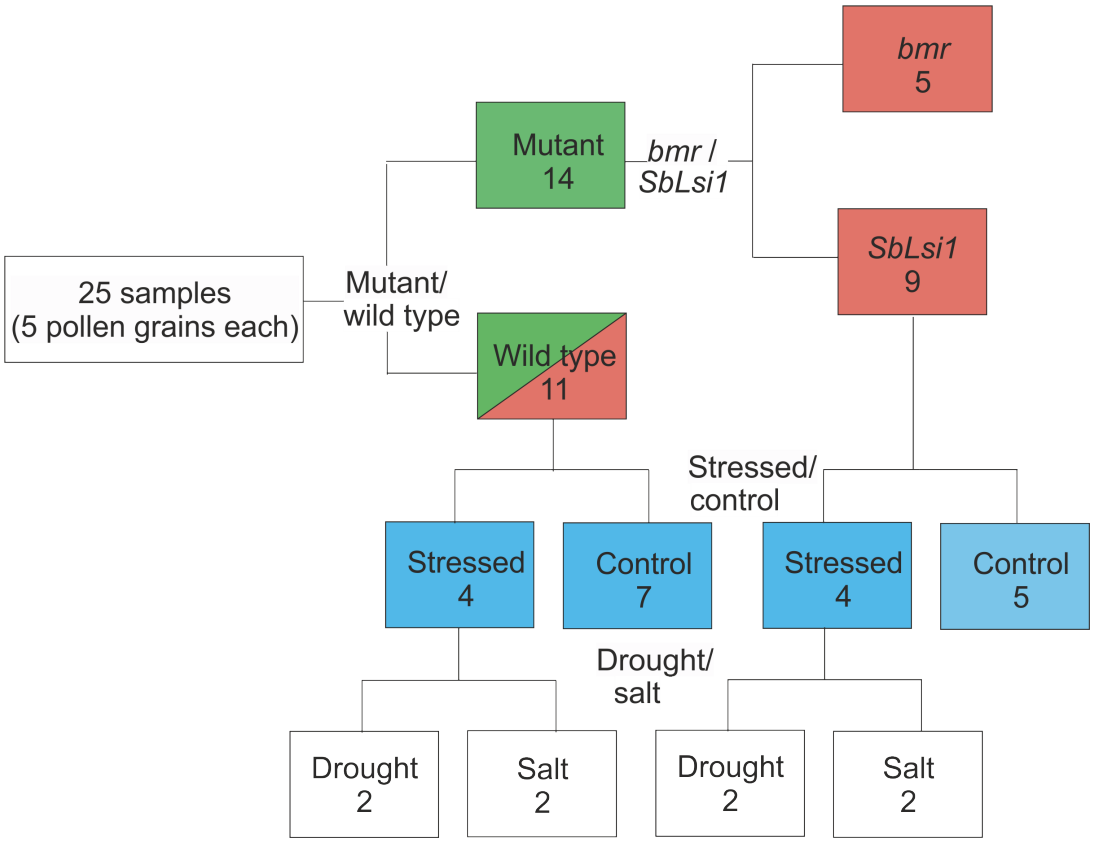

Figure 6.9: Schematic presentation of the numbers of samples from the sample set Pollen Israel. The schematic also indicates the subsections in which a specific classification problem is discussed: The discrimination of pollen from wild-type and both mutants as one class (green), and from wild-type and both kinds of mutants (*bmr* and *SbLsi1*, red), as well as regarding the separation between different growth conditions, here initiated by stress (drought and salt together, blue).

For each of the 25 samples, 5 different pollen grains were measured using Raman microspectroscopy. 10×10 spectra were obtained, that cover $20\ \mu m \times 20\ \mu m$ of the pollen grains, to cover also variances within the ultra-structure of the pollen grain.
The quality of the obtained spectra can vary greatly, e.g., due to variation in signal-to-noise, sample autofluorescence, or indication of radiation damage. In order to analyze small effects on the pollen spectra, induced by, e.g., growth conditions of the parental plant, such spectra could be treated as outliers and eliminated from the analysis. However, exclusion of possibly large portions of sample sets from the experiment cannot be a solution, specifically if many spectra are concerned as in the case of strong autofluorescence of a particular pollen grain. In some cases, such 'quality parameters' can be interpreted as real biological differences between groups of samples.
In this subsection, the influence of non-Raman information on classification and spectral characterization will be investigated using averages of 125 mapping data sets, corresponding to all probed pollen grains from the respective two sample groups (wild-type, mutants) (indicated in Figure 6.9, red).

Figure 6.10 shows the PCA results of the 125 averaged raw spectra colored by their affiliation into mutants (black) and wild-type (red). There is no discrimination between the raw spectra between pollen from mutant and wild-type plants. PC 1 shows a strong variation in the fluorescence background with an explained variance of 99.4 %. Too many spectra have the problem of a strong underlying fluorescence. To have an appropriate amount of spectra, they cannot be excluded from the rather small data set. For suitable data analysis, the data need to be pre-processed carefully.

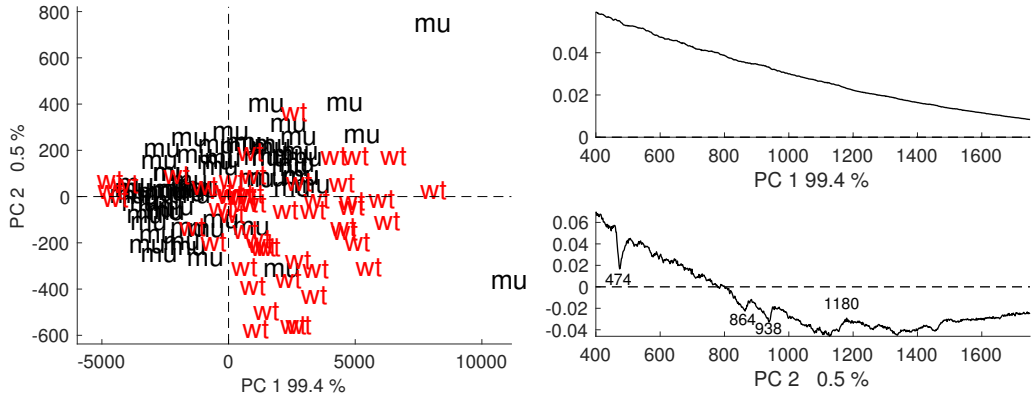

Figure 6.10: (Left) Scores plot and **(right)** loadings for the PCA of 125 averaged raw spectra from the 125 pollen grains. Coloring refers to separation regarding mutants, (*SbLsi1* and *bmr*, black) and wild-type (red) (compare with Figure 6.9, green).

Figure 6.11 (left) shows all 100 spectra from one map of a pollen grain from a *bmr*-mutant plant with high fluorescence (black spectra) and the 100 spectra from one with low background issues (red spectra), here from a wild-type plant. It has to be pointed out that the background issue is not specific for the discrimination of wild-type and mutant spectra, as also indicated in the scores plot in Figure 6.10. In the spectra with lower background, several bands become visible, while in the black spectra, Raman signals cannot be seen. In order to avoid radiation damage (see previous subsection), spectra were obtained using an acquisition time of 1 s. Many suggestions for fluorescence correction can be found in the literature, for example, by the experiment, e.g., photo destruction[13] or by using a time-gated detection[213,214] as well as mathematically. Commonly a baseline is estimated and subtracted from the spectra.[172,173] Here, the effect of a mathematical baseline correction is assessed. Therefore, Eiler's asymmetric least square (AsLS) algorithm was applied.[177,178] In Figure 6.11 (right), the averaged spectra for both maps are shown in black. The AsLS - correction iteratively estimates the baseline, to minimize the residuals. The estimated baselines are drawn in red in Figure 6.11. The baselines are subtracted from the raw spectra afterwards, to result in spectra where background issues are reduced (not shown here).

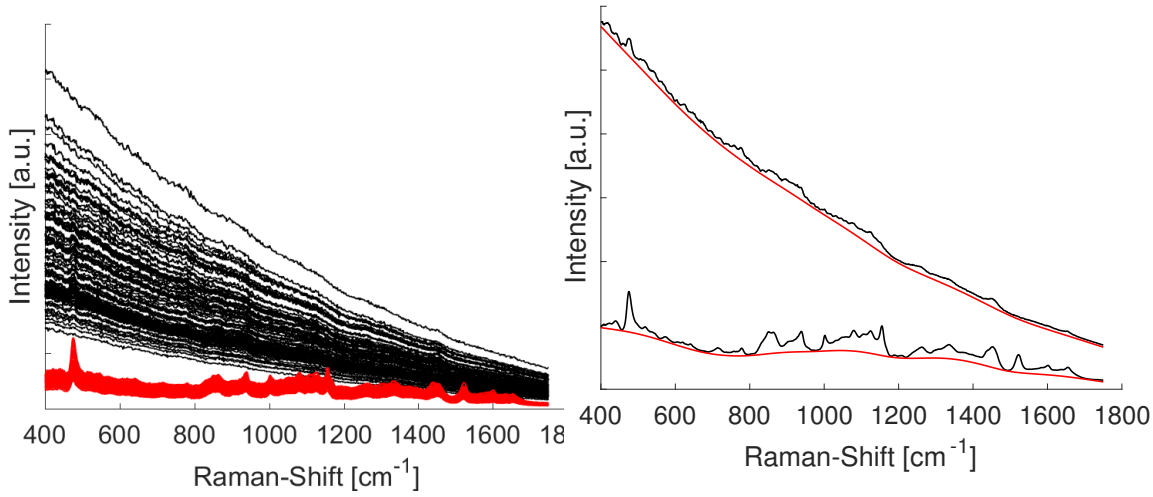

Figure 6.11: (Left) Raw spectra from one map with strong fluorescence background issues (black) and one map with low background issues (red). (Right) Averaged spectrum in black and corresponding AsLS-baseline (red) from the two maps is shown on the left side.

In order to eliminate effects due to variation in absolute intensity, the spectra should ideally be normalized. Figure 6.12 shows the same set of spectra as presented in Figure 6.11 after AsLS baseline correction and vector normalization. In the case of low fluorescence, bands can be detected and assigned well. Spectra from maps with high fluorescence background show few and broad bands with a low signal-to-noise-ratio. Several bands occur at the same Raman-Shift. Most prominent are the bands at 475, 781, 938, 1126, and 1440 cm^{-1}. These bands can be tentatively assigned to starch and proteins.[5,10]

The averaged spectra of both maps in Figure 6.12, right show a better signal-to-noise-ratio. Thus the smaller bands e.g. at 641, 778, 1261, 1337, 1602, and 1654 cm^{-1} are more pronounced in both averaged spectra. After the spectral pre-processing, both spectra look very similar. Even so, a few differences can still be seen by eye, most importantly the bands at 1002, 1152, and 1522 cm^{-1}, assigned to vibrations of carotenoids.[5,13] The pre-processing using AsLS baseline correction and vector normalization is now applied to all mapping data before calculation of the 125 average spectra that correspond to the respective 125 pollen grains (see also Figure 6.10) and the PCA is repeated (Figure 6.13).

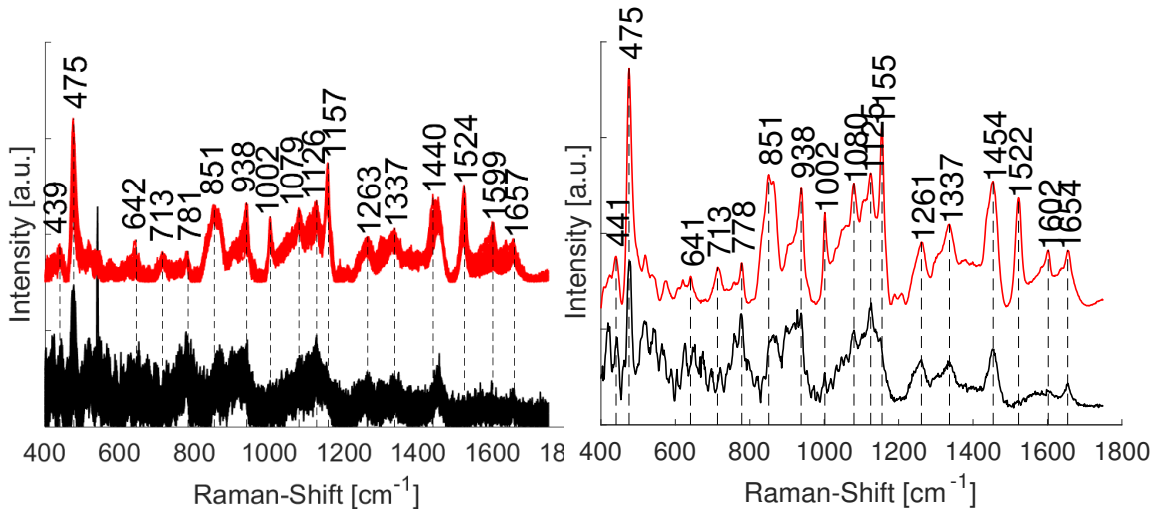

Figure 6.12: (Left) Baseline and vector normalized spectra from one map with strong fluorescence background (black(see also black spectra in Figure 6.11) and one map with low fluorescence (red). **(Right)** Averaged spectrum from the two maps is shown on the left side. Spectra are stacked for clarity.

6.3 Assessment of within-species-variation in pollen grains using Raman microscopy

The scores plots in Figure 6.13 (A, B, and C) presents the distribution of the score values after the PCA of the spectra from the 125 pollen grains for the three different classification problem, namely the discrimination of wild-type and mutant plants in Figure 6.13 (A) (see Figure 6.9, green), wild-type and the two different mutants in Figure 6.13 (B)(see Figure 6.9, green), and the different growth conditions (see Figure 6.9, blue) in Figure 6.13 (C). It differs greatly from the scores plot of the same data set when PCA is conducted without pre-processing, indicating that non-Raman features, specifically the baseline due to fluorescence can play an important role in classification outcome. The spectra of pollen grains from wild-type plants have positive values and the spectra of pollen grains from mutant plants have negative values for PC 2 (Figure 6.13 (A), black), the spectra can be differentiated. According to the corresponding loadings in Figure 6.13 (D), the spectra differ among others in their signals at 473, 854, and 1412 cm^{-1}. These bands can be assigned to starch and proteins, respectively.[5,10,31] The loadings of PC 1 and PC 2 indicate that the relative contributions of carbohydrates such as starch and of proteins differ in the pollen of the two mutants and the wild-type plants. In Figure 6.13 (B), the same scores plot from Figure 6.13 (A) is colored with respect to the wild-type and the two different mutants. Particularly, the score values of the *bmr*-mutant (Figure 6.13 (B), black) show no separation. The coloring of the score values regarding the different stress conditions (Figure 6.13 (C)) indicates that also the variance between the different growth conditions cannot be assessed using this data set here.

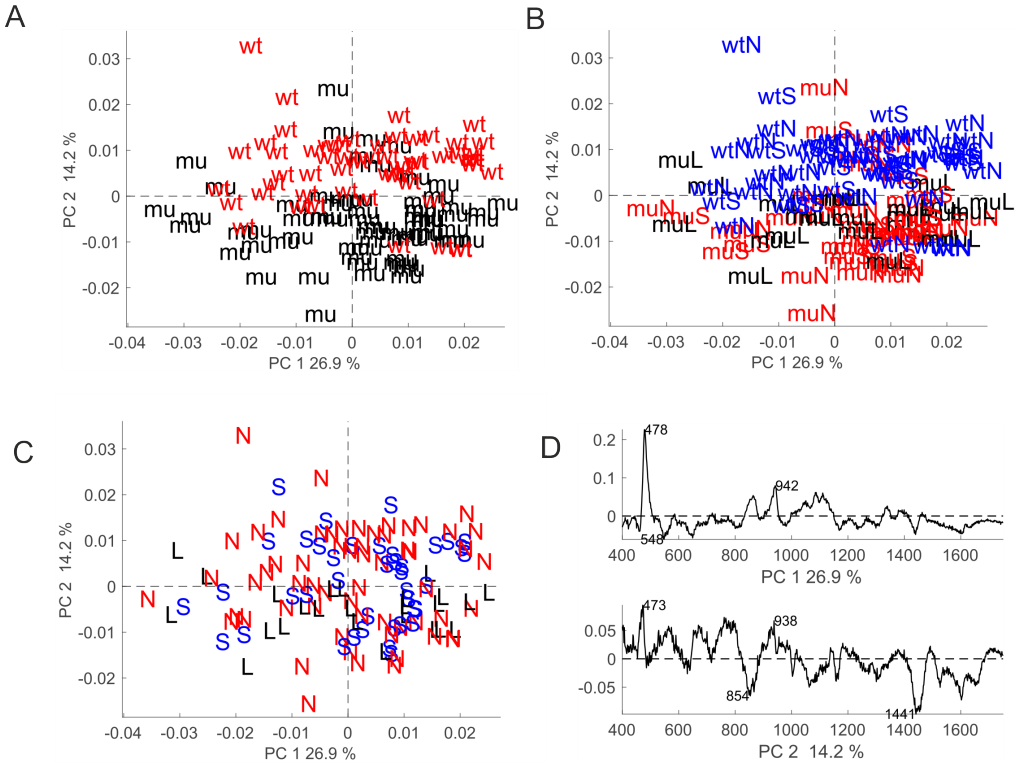

Figure 6.13: (**A**, **B**, and **C**) Scores plots and **(D)** loadings for the PCA of 125 averaged spectra from the 125 pollen grains. Coloring refers to **(A)** mutant, black and wild-type, red. **(B)** *SbsLi1*-mutant, red, *bmr*-mutant, black, and wild-type, blue. **(C)** The different stress conditions and the *bmr*-mutant.(*bmr* mutant, L, black; stressed plants (drought and salt stress) together, S,blue; control groups, N, red). Spectra were pre-proccesed using AsLS baseline correction and vector normalization before averaging.

6.3.1 Combination of Raman and MALDI MS data to assess within-species-variation in pollen

A subset of the pollen samples was also investigated using MALDI experiments. The quality of the MALDI spectra can also vary enormously for the pollen samples, respectively. In contrast to the Raman measurements, mass spectra with very low signal-to-noise-ratio are discarded by the software in MALDI-TOF MS experiments. Therefore, the MALDI-subset contains 14 spectra including eight pollen samples from eight mutant plants and six samples from wild-type plants.

Figure 6.14 shows the results of the PCA on the 14 spectra. The six spectra from pollen samples of wild-type show mostly negative score values regarding PC 3 (Figure 6.14, left, blue and green markers). In addition, most spectra from samples of mutants have positive values for PC 3 (Figure 6.14, left, red and black markers).

However, the main variance of the data set is based on the variation between the two different breeding times of the plants (A, Figure 6.14 (left), blue and black markers, and B, Figure 6.14 (left), red and green markers). The impact of different breeding times of the plants can also

be detected in Raman data but is, among the other effects described above, minimized by pre-processing. As discussed in Chapter 4 and Chapter 5, MALDI is sensitive towards variation based on population or genotypes of plants, whereas Raman spectroscopy is a more powerful tool to detect variations in growth conditions. Therefore, in MALDI discrimination of the mutant/wild-type is in high competition with the different breeding times. In the case of the discrimination between mutant and wild-type plants, Raman spectroscopy would lead to better discrimination of the spectra.

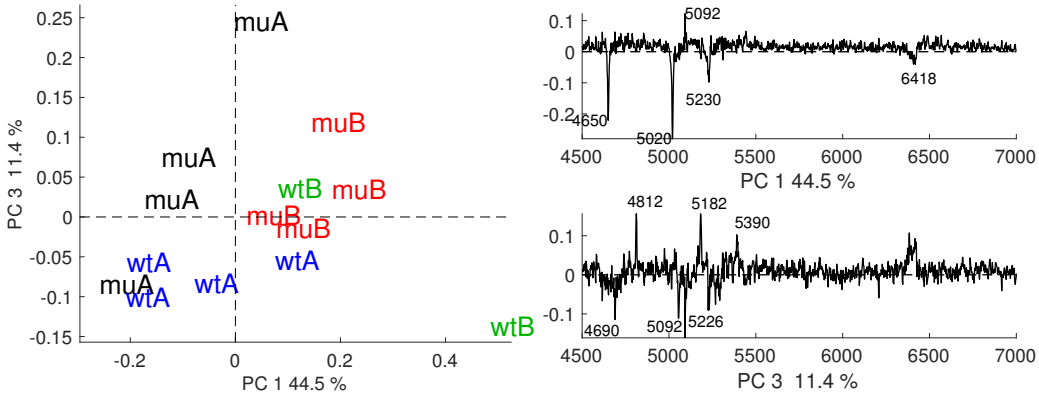

Figure 6.14: (Left) Scores plot and **(right)** loadings for the PCA of 14 mass spectra. Coloring refers to mutant, black and wild-type, blue from breeding time A, as well as mutant, red and wild-type, green from breeding time B. Spectra were pre-processed using AsLS baseline correction and vector normalization.

Consensus principal component analysis (CPCA) can be applied to Raman and MALDI MS data to find a common pattern within the data. Therefore, the corresponding Raman subset of 14 samples was taken from the Raman data set to enable an equal number of samples, which is required for CPCA. Figure 6.15 and Figure 6.16 show the global and block scores for the CPCA of the Raman and MALDI data with a coloring regarding the mutant and wild-type conditions. The global scores indicate a separation of the mutant and wild-type spectra as seen in Figure 6.15. Global score values from mutants have mostly positive values regarding PC 1 and PC 2 (Figure 6.15, black), In comparison, the scores from wild-type spectra have almost all (5 out of 6) negative values for PC 2 (Figure 6.15, red). Furthermore, the score values show in the first and second CPC a smaller in-group variation compared to the wild-type ones (Figure 6.15, red). The explained variances of the shown PCs are 31 % and 22 % respectively.

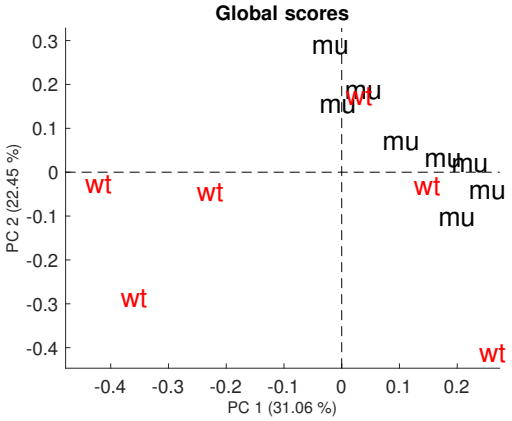

Figure 6.15: Global scores plot for the CPCA of 14 Raman spectra and 14 mass spectra. Coloring refers to mutant, black and wild-type, red.

The explained variances of CPC 1 of the block scores are 47 % for Raman block and 15 % for the MALDI block (Figure 6.16). This indicates that the Raman block has a higher influence on the global scores (Figure 6.15) regarding CPC 1. Score values from mutant spectra have all positive values for CPC 1, whereas most of the score values of the wild-type spectra are negative. The first CPC in the Raman Block separates score values from mutant and wild-type spectra(Figure 6.16 (left)). On the contrary, in the MALDI block the mutants score values are mostly positive for CPC 2 (Figure 6.16 (right), black). Score values for the wild-type spectra are mainly negative for both the Raman CPC 2 (Figure 6.16 (left), red) and the MALDI block CPC 2 (Figure 6.16 (right), red).

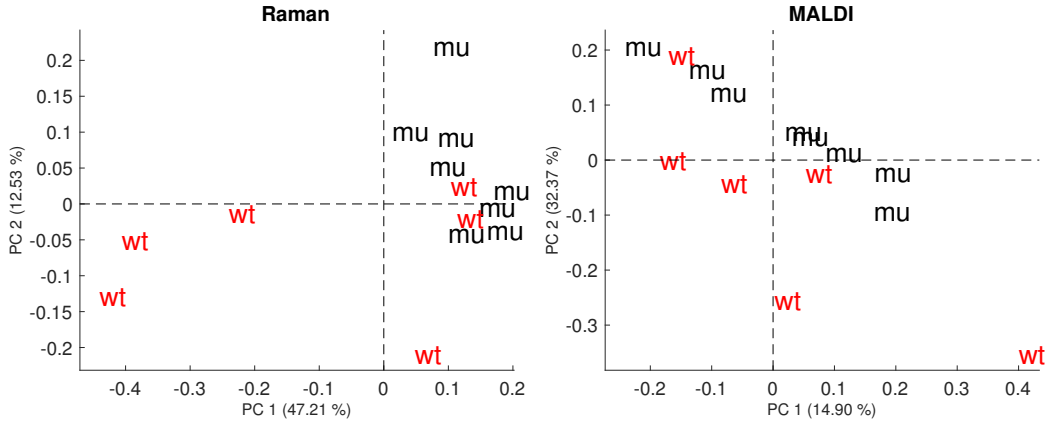

Figure 6.16: (Left) Raman and **(right)** MALDI block scores plots for the CPCA of 14 Raman spectra and 14 mass spectra. Coloring refers to mutant, black and wild-type, red.

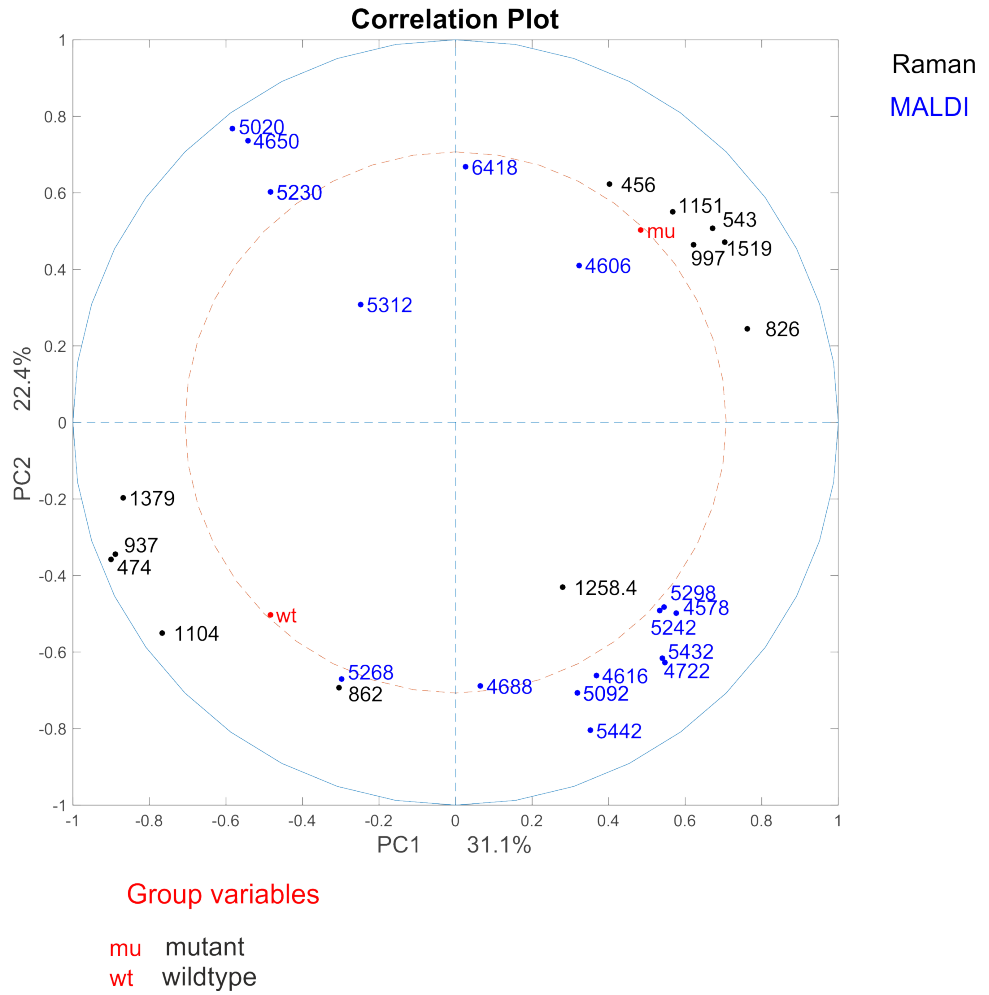

Figure 6.17: CPCA correlation loadings plot for the first and second CPC. Displayed are the global scores of the two kinds mutant (red, mu) and wild-type (red, wt), as well as the loadings of the blocks of Raman (black) and MALDI-TOF (blue). For clarity only extrema of the loadings were shown for the spectroscopic/ spectrometric data.

The CPCA enables the investigation of a common underlying pattern. The loading values of both, Raman and MALDI can be represented in a correlation loadings plot. Figure 6.17 shows the correlation loading plot after the CPCA using 14 Raman and MALDI spectra. The values of the global scores for mutant and wild-type spectra are drawn in red, the Raman loadings in black, and the MALDI loadings in blue. For comparison, just extrema of the loadings are shown. Values that are more correlated with each other would have a similar position in the scores plot and *vice versa*. Close to the mutant score values, several Raman signals are grouping together, namely the signals at 456, 543, 826, 997, 1151, and 1519 cm^{-1}. The last three bands can be assigned to carotenoids.[13]

Regarding the MALDI signals, the peak at *m/z* 4606 can be correlated with the scores of the mutants. In addition, the MALDI peaks at *m/z* 4650, 5020, 5312, and 6418 are corresponding to the separation regarding CPC2. Furthermore, the Raman bands at 474, 937, 1104, and 1379 cm^{-1} are close to the centroid of the wild-type spectra. These Raman bands, that can

mainly be assigned to carbohydrates can also be associated with the separation of the wild-type spectra. The CPCA yields a separation of mutant and wild-type spectra using CPC 1 and CPC 2. In the following discussion, the higher CPCs are investigated in order to find a common pattern for the discrimination of plants that suffer stress from the control group. In the individual PCA of the Raman spectra, a separation regarding the different growth conditions could not be achieved (Figure 6.13 (C)). Here, a subset of 14 spectra was analyzed using Raman and MALDI spectra in a CPCA.
Figure 6.18 shows the global scores of CPC 3 and CPC 4 from the same multiblock analysis shown above (compare with Figure 6.15). It can be seen that most score values of the spectra from the control group have negative values regarding CPC 4. Nevertheless, a separation of the score values cannot be detected. Particularly, five out of seven spectra show positive score values regarding CPC 3 and five score values of spectra from stressed plants group together mostly having negative values regarding CPC 3.

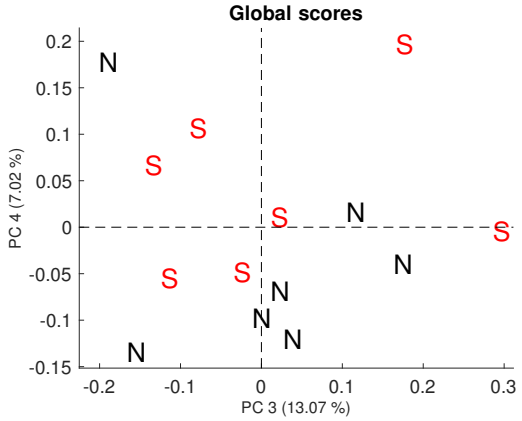

Figure 6.18: Global scores plot for the CPCA of 14 Raman spectra and 14 mass spectra. Coloring refers to control, N, black and stressed, S, red.

Also in the block scores, no efficient separation can be found, but they indicate a high grouping of the control group in the MALDI scores regarding CPC 4. Therefore, it is suitable to check the correlation loading plot regarding the grouping of spectra from the control group (Figure 6.20).
As discussed above, the centroids are close to 0 regarding the third CPC, which indicates that the separation is not well explained by this component. In addition, the two centroids are in the inner circle which corresponds to an explained variance below 50 %. The MALDI peaks responsible for the dragging of the score values from the stressed plants are at *m/z* 4812, 5180, 5392, 5334, and 6388. On the opposite side with negative values, score values from the control group are dragged towards peaks at *m/z* 4686, 5096, 5060, 5348, and 5458.
The CPCA indicates a low variation of pollen from plants that recover from drought and salt stress with regard to the control group. The variation in pollen is higher in the case of a mutant/wild-type separation. Drought and salinity stress is well-known to decrease

pollen quality enormously. With CPCA no conclusion about a connection between pollen quality can be made. Reasons might be a masking from the mutant/wild-type separation.

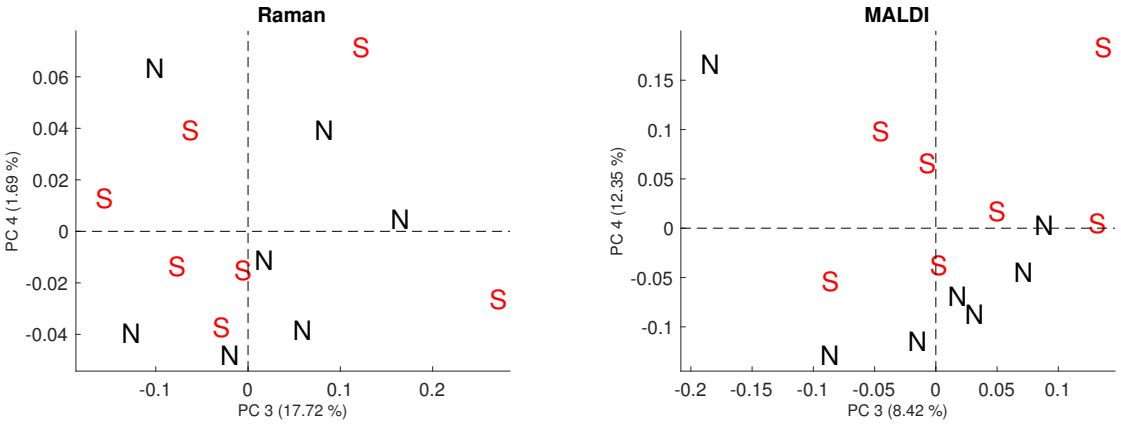

Figure 6.19: (Left) Raman and **(right)** MALDI block scores plots for the CPCA of 14 Raman spectra and 14 mass spectra. Coloring refers to control, N, black and stressed, S, red.

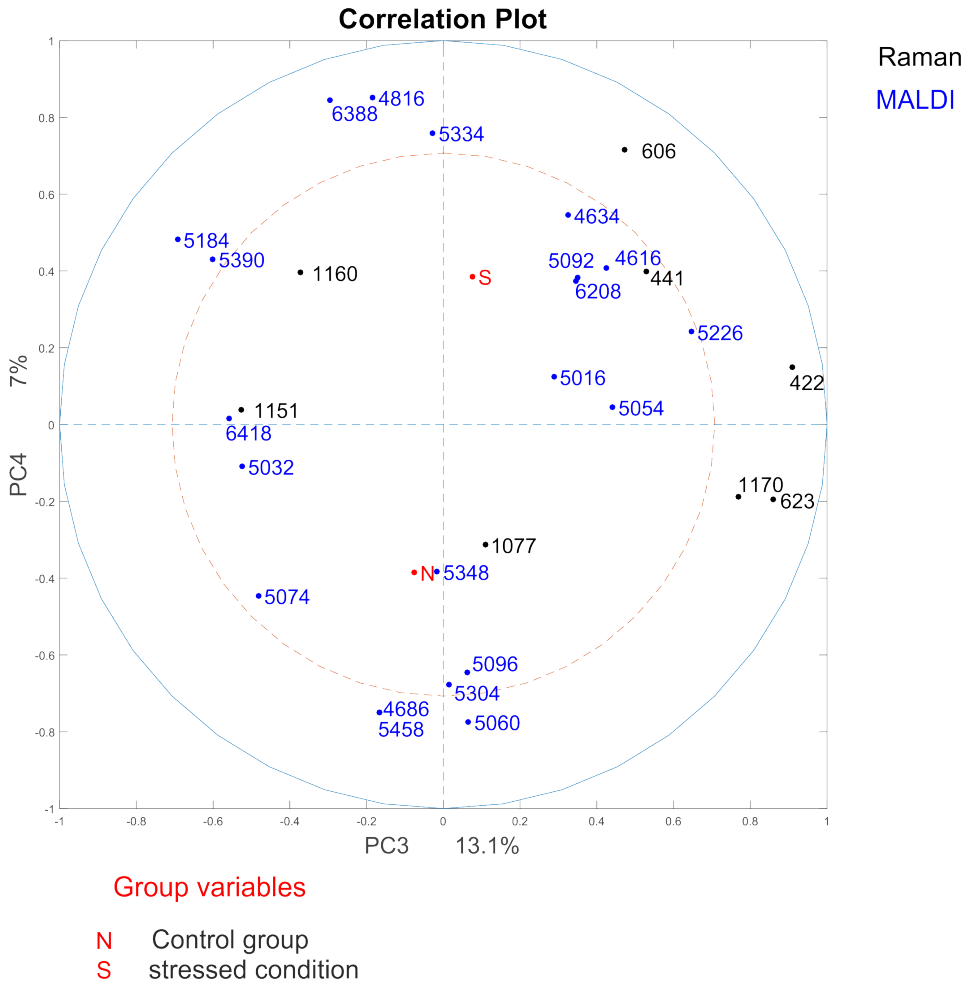

Figure 6.20: CPCA correlation loadings plot for the 1st and 2nd CPCA component. Displayed are the global scores of the two groups recovering from stress (red S) and the control group (red N), as well as the loadings of the blocks of Raman (black) and MALDI-TOF (blue). For clarity only extrema of the loadings were shown for the spectroscopic/ spectrometric data.

This chapter addressed various problems that arise during the acquisition and analysis of Raman mapping data from pollen grains. Raman data obtained from two different substrates were compared. It was shown that Raman spectra of pollen grains on calcium fluoride do not differ greatly from spectra of pollen grains fixated on carbon tape. Therefore, fixation on carbon tape could be pursued in future experiments with biological samples, as it provides additional advantages regarding multimodal analysis over direct measurements on calcium fluoride without fixation.

Data from unintentional germinated pollen grains were discussed, as a common problem that can occur in the sampling of pollen.Since single pollen grains can be measured using Raman microspectroscopy, such effects, as well as damage of grains need to be considered. An optimized pre-processing can help to analyze the pollen samples as well. In addition, the characterization of variances in a complex structure from samples of the same species *Sorghum bicolor* was conducted, by combining pre-processed Raman spectra and MALDI

TOF MS.The spectra set consists of variation in mutant and wild-type as well as stressed and not stressed conditions of the parent plant. Multiblock analysis consisting of Raman and MALDI spectra can help to discriminate between mutants and wild-type, but cannot differentiate between different stress conditions here.

7 Discrimination and characterization of different grass pollen species using FTIR spectra of single pollen grains

Parts of the results presented in this chapter are published in: Diehn, S., Zimmermann, B., Tafintseva, V., Bağcıuğlu, M., Kohler A., Ohlson M., Fjellheim S, and Kneipp J.., Discrimination of grass pollen of different species by FTIR spectroscopy of individual pollen grains. Anal Bioanal Chem (2020).
https://doi.org/10.1007/s00216-020-02628-2

Fourier transform infrared (FTIR) spectroscopy in combination with chemometric tools gives accurate results in discriminating different pollen species. Still, most of the studies include extraction steps or a high amount of sample material. Unfortunately, because of small grain size, pollen grains cause scattering artifacts in FTIR spectra. Before measurements of single pollen grains, the samples need to be embedded in paraffin. Thereby scatter effects are reduced, due to the optical properties of paraffin.[14] Once valuable information about the chemical composition of pollen can be obtained from the FTIR spectra, machine learning approaches, as well as consensus principal component analysis (CPCA), can be applied for a comprehensive analysis of pollen composition.
The reproducibility of FTIR of embedded pollen grains is evaluated using the same sample set as in the prior analysis (Pollen Norway IIa) and a new sample set in addition (Pollen Norway IIb). Figure 7.1 gives an overview of the samples used in this chapter.

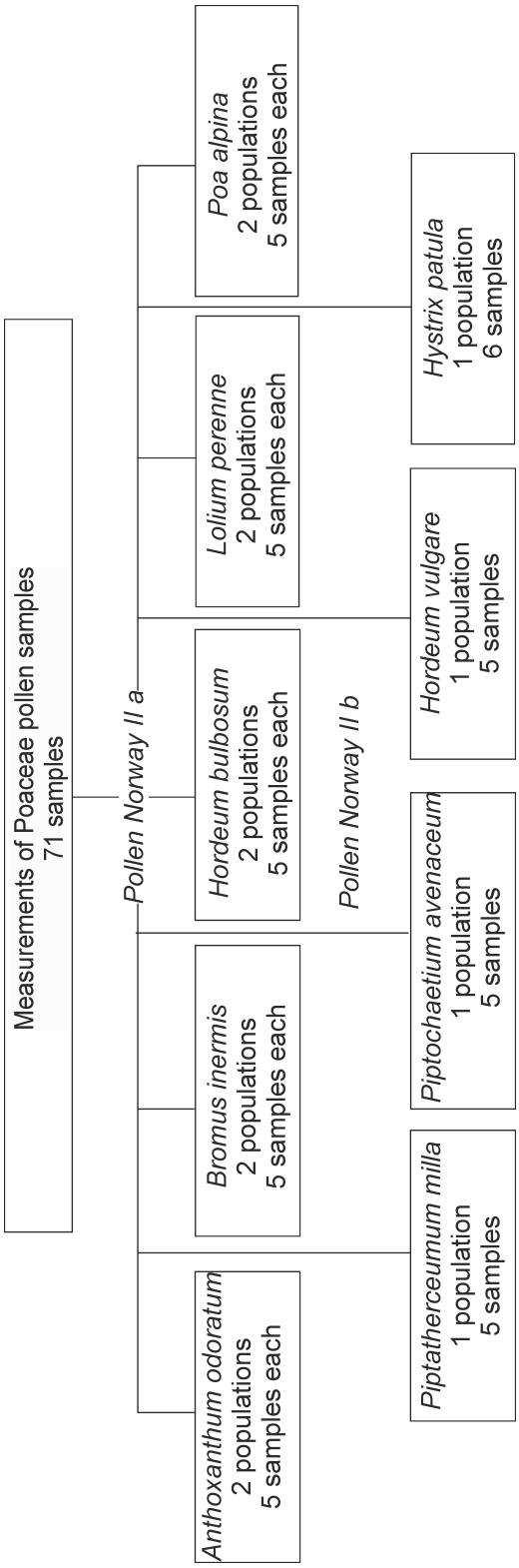

Figure 7.1: Schematic overview of the sample set Pollen Norway II. Pollen Norway IIa consists of the five pollen species *Poa alpina*, *Anthoxanthum odoratum*, *Lolium perenne*, *Bromus inermis* and *Hordeum bulbosum*. Pollen Norway IIb comprises the species *Hystrix patula*, *Hordeum vulgare*, *Piptatherum millaceum*, and *Piptochaetium avenaceum*.

7.1 Evaluation of FTIR spectra from embedded and non-embedded pollen grains

The bright-field images (Figure 7.2) display the dry pollen grains from the five different grass species are similar in size and morphology. In general, grass pollen grains are characterized by simple spherical shape and microechinate grain wall ornamentation.[215] Grass pollen has very limited mechanisms for preventing desiccation.[216] As a result, grass pollen morphology is dramatically changed after shedding, collapsing from spherical shape of fresh pollen to extensive infolding of dry pollen.[217] It leads to a large variation in Mie-scattering effects, resulting in extremely unreproducible spectra. Although the pollen grains of all five measured species have similar morphology, those of *Poa alpina* and *Anthoxanthum odoratum* are slightly smaller than the pollen grains of *Lolium perenne*, *Bromus inermis*, and *Hordeum bulbosum*.

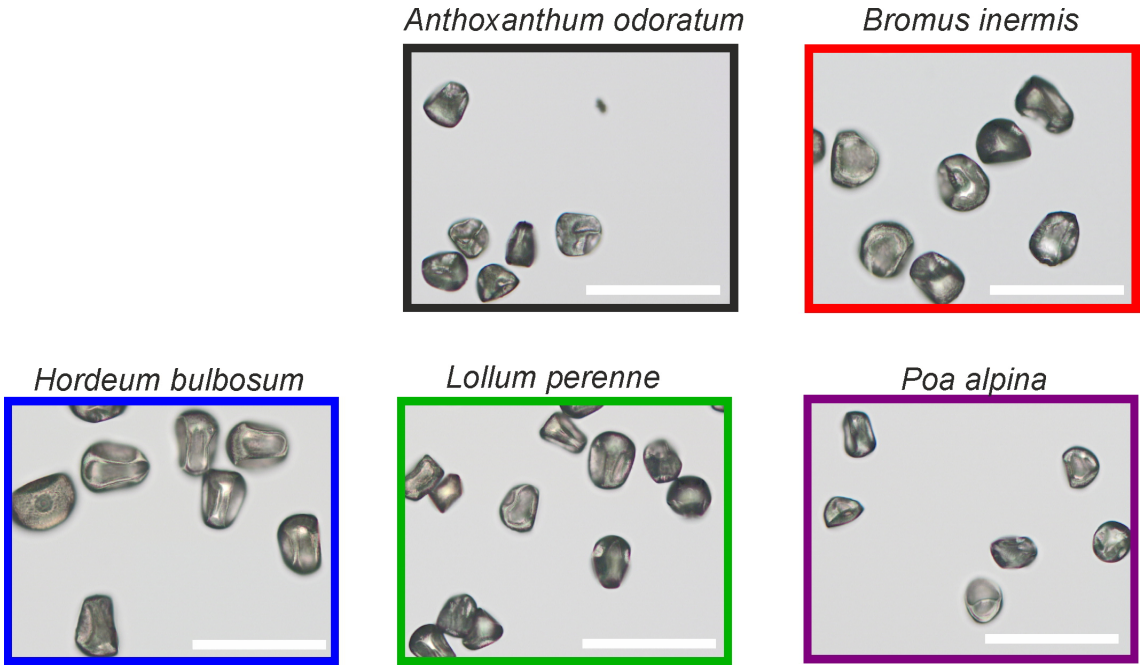

Figure 7.2: Bright field images of dry pollen grains from the 5 different grass species *Poa alpina*, *Anthoxanthum odoratum*, *Lolium perenne*, *Bromus inermis* and *Hordeum bulbosum*. Scale bar, 100 μm.

Following an established protocol presented by Zimmermann *et. al.*,[14] the pollen samples were embedded in paraffin to avoid scattering artifacts in the spectra. Figure 7.3 shows representative raw spectrum of one pollen grain on ZnSe (Figure 7.3, blue), a pollen grain on ZnSe embedded in paraffin (Figure 7.3, red) and of the paraffin layer without any pollen grains on ZnSe (Figure 7.3, black) in the spectral range from 650 - 3800 cm^{-1}. The influence of the paraffin layer is particularly pronounced in the spectral regions at 1300- 1500 cm^{-1} and 2800- 3100 cm^{-1}. In the spectral region 1300- 1500 cm^{-1} two bands occur at 1377 cm^{-1}

and 1462 cm^{-1}. These bands can be assigned to the methyl rocking vibration at 1377 cm^{-1}, the CH_2 bending, and CH_3 deformations modes at 1462 cm^{-1}.[218] The spectral region 2800-3100 cm^{-1} is saturated and associated with CH stretching.[218]

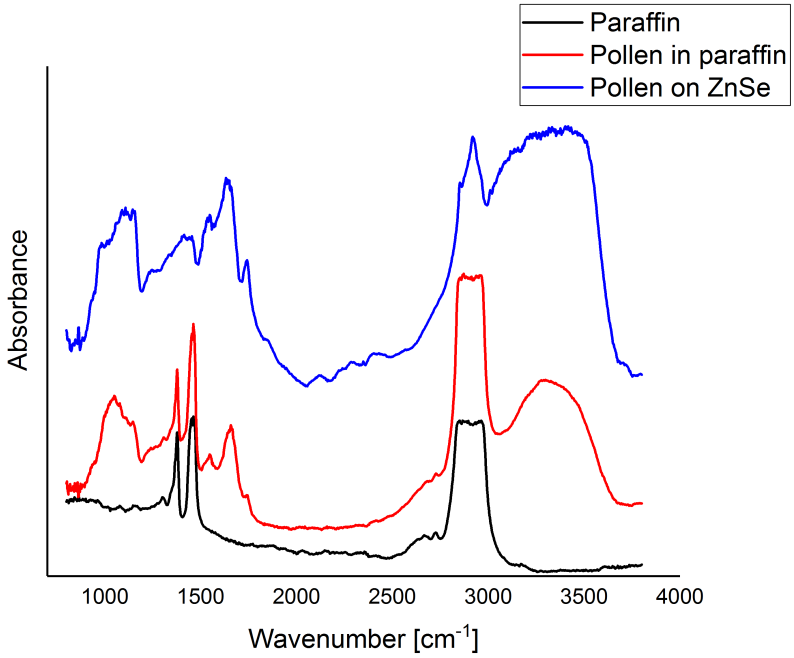

Figure 7.3: Representative raw FTIR spectra of a pollen grain on ZnSe (blue), a pollen grain on ZnSe embedded in paraffin (red) and of the paraffin layer (black) in the spectral range from 650 - 3800 cm^{-1}.

For biotyping and characterization of pollen samples, the spectral region between 800 and 2000 cm^{-1} is more useful than the CH and OH stretching region above 2000 cm^{-1}. The two additional paraffin bands at 1377 and 1462 cm^{-1} interfere with bands from the pollen spectrum (Figure 7.3). In comparison, the spectrum of embedded pollen (Figure 7.3, red) shows less scattering artifacts than the spectrum of non-embedded pollen grains (Figure 7.3, blue). In particular, the region between 2000 and 2500 cm^{-1} where no strong signals are expected for pollen spectra[14] is an indication for the scatter effect in the non-embedded pollen (Figure 7.3, blue).

To compare spectra from non-embedded and paraffin-embedded samples, the spectra were pre-processed as follows: The spectra were baseline-corrected using asymmetric least squares (AsLS) correction, as proposed by Eilers[177,178] and vector-normalized before averaging. In Figure 7.4 the average spectra of non-embedded (Figure 7.4, A) and of the paraffin-embedded pollen grains (Figure 7.4, B) for each pollen species are presented. The spectra of the embedded pollen show much less variation within each species (Figure 7.4, B) compared to the large

standard deviation of the unembedded samples (Figure 7.4, A). The most prominent bands in the spectra are found at 989, 1045, 1161, 1549, 1659, and 1745 cm^{-1}. These bands can be assigned to carbohydrates (989 and 1045 cm^{-1}), lipids (1161 and 1745 cm^{-1}), and amide II and amide I vibrations of proteins (1549 and 1659 cm^{-1}).[5,10,80] Besides, the characteristic absorbance of paraffin adds to this pollen signature and is particularly prominent in the region from 1300 - 1500 cm^{-1} (Figure 7.4, B).

The paraffin bands at 1377 and 1462 cm^{-1} in the spectra of the embedded samples vary between the different species (Figure 7.4, B). In the spectra of pollen from *Poa alpina* and *Anthoxanthum odoratum*, both bands have higher relative absorbance values, compared to *Lolium perenne* and *Bromus inermis*. In the spectrum of *Hordeum bulbosum*, the two characteristic paraffin signals have smaller contributions and the spectrum in the region from 1300-1500 cm^{-1} resembles that of the averaged spectrum from the non-embedded pollen grains (Figure 7.4, left and right). The different relative contributions of the embedding paraffin in the spectra of the three species are likely related to the sizes of the pollen grains. This leads to a different relative amount of paraffin vs. pollen material in the probed microscopic volume.

Four different approaches for correction of FTIR spectra of the paraffin-embedded samples will be discussed using principal component analysis (PCA) and partial least-square discriminant analysis (PLS-DA).

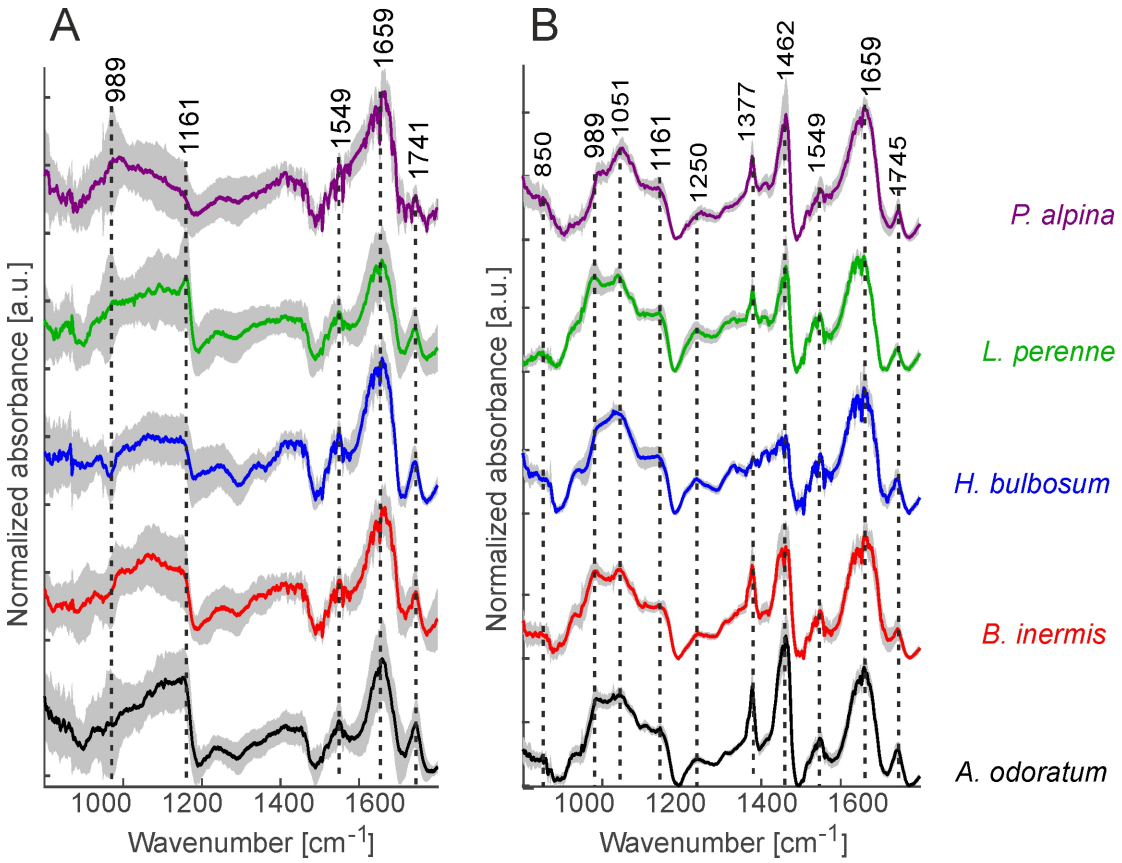

Figure 7.4: Pre-processed and averaged spectra from **(A)** non-embedded pollen grains and **(B)** paraffin embedded pollen grains for the five different grass species *Poa alpina*, *Anthoxanthum odoratum*, *Lolium perenne*, *Bromus inermis* and *Hordeum bulbosum*.

7.2 Utilization of FTIR spectra of embedded single pollen grains

The first approach deals with spectra, that are extended multiplicative signal correction (EMSC) corrected without any further consideration about the paraffin. In a second approach, the paraffin affected spectral region from 1300-1500 cm^{-1} is omitted and the two spectral regions from 800-1300 cm^{-1} and 1500-1800 cm^{-1} are either concatenated or combined in a multiblock analysis. Non-negative matrix-factorization (NMF) was applied in a third approach to split each spectrum into a pure paraffin and a pure pollen spectra component. Furthermore, a modified EMSC algorithm using a paraffin spectrum as a constituent was applied as a fourth approach to minimize the variances from the paraffin contribution within the spectra.

Figure 7.5 shows an overview of the four different approaches and how to deal with a paraffin contribution, namely approach 1) without further consideration, approach 2) omitting the paraffin affected spectral region, approach 3) extraction of pure pollen and pure paraf-

fin contribution using NMF, and approach 4) EMSC using a paraffin constituent spectrum. Figure 7.5 outlines the data processing steps that include basic pre-processing such as baseline correction, normalization, and calculation of average spectra, as well as the steps that were used specifically to assess the contributions by paraffin to the spectra. The spectral region of 800 to 1800 cm^{-1} was selected since it contains bands that are distinctive for pollen grains.[5,14,118]

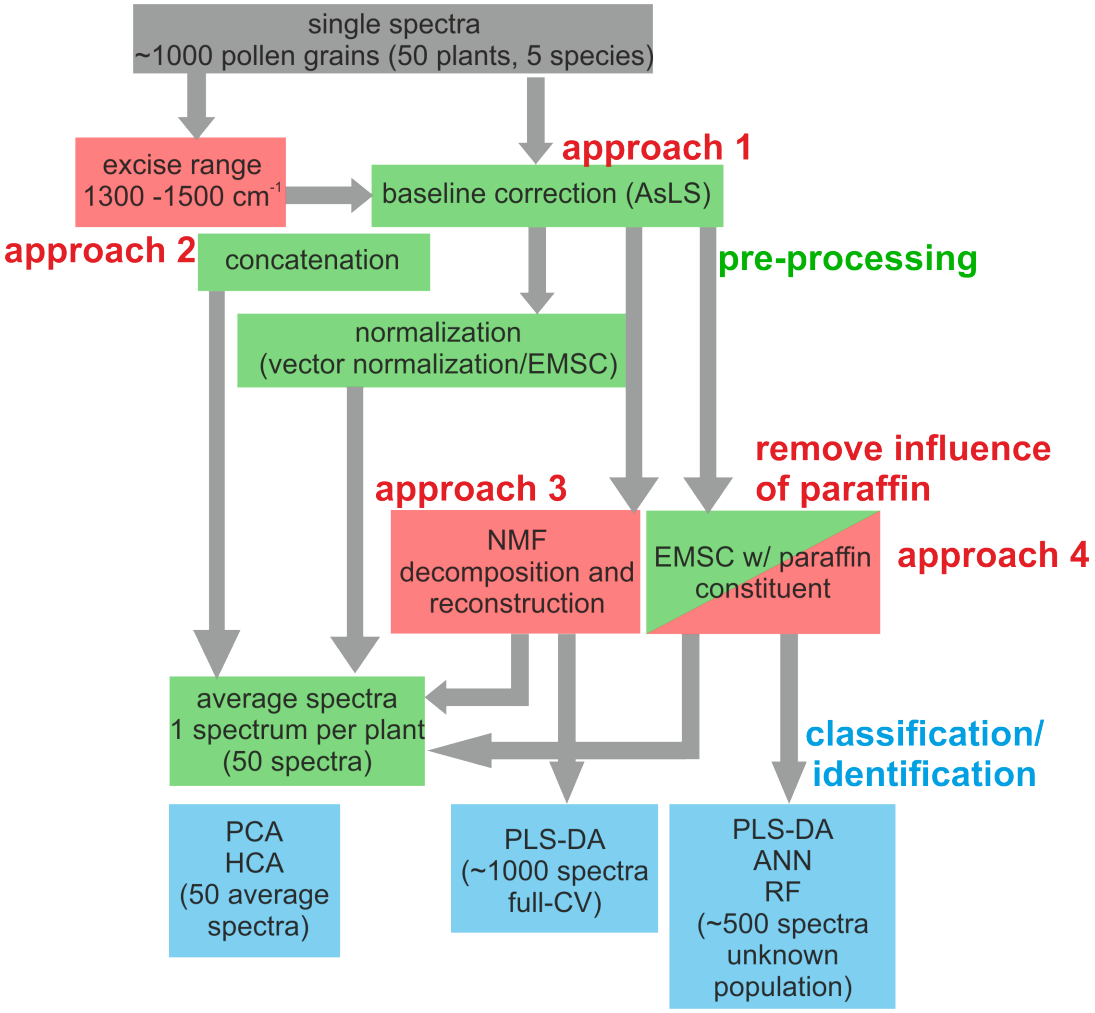

Figure 7.5: Schematic representation of the 4 different approaches and their specific pre-processing.

7.2.1 Approach 1: without further consideration of the paraffin contribution

The spectra from paraffin-embedded pollen grains are baseline corrected by AsLS before applying a simple EMSC, a MSC model extended by a linear and quadratic component[131] that replaces normalization. Subsequently, the spectra were smoothed using a Savitzky-Golay filter with a window size of 9 (see Chapter 3.3.1) and a second-order polynomial.[174] For

classification by PLS-DA, the individual spectra were used. For analysis by hierarchical cluster analysis (HCA) and PCA, averages of the spectra of each plant were calculated.
The assessment of this pre-processing by PLS-DA with leave-one-out-cross-validation (full-CV) indicates that the spectra of the different species can be discriminated (Table 7.1). An overall success rate of 79 % was achieved, while the individual success rates reached a value of approximately 90 % for *Hordeum bulbosum*, *Anthoxanthum odoratum*, and *Poa alpina* spectra.

Table 7.1: Identification of 1004 pollen spectra according to their species with PLS-DA with 9 latent variables and leave-one-out-validation. Spectra were pre-processed using approach 1.

affiliation / identified by PLS-DA as	*A. odoratum*	*B. inermis*	*H. bulbasum*	*L. perenne*	*P. alpina*
A. odoratum	184	3	5	13	11
B. inermis	1	114	10	14	2
H. bulbasum	4	51	189	5	2
L. perenne	5	18	0	131	7
P. alpina	5	14	5	33	178
success rates	92 %	57 %	90 %	67 %	89 %
Overall success rate	79 %				

The average spectra in Figure 7.4 already suggest that the paraffin spectral contribution could influence the discrimination of the different pollen species. The results of the PCA corroborate this, and the loadings of the first principal component (PC 1) (Figure 7.6, A, right) show the two strong paraffin-related signals at 1377 and 1462 cm^{-1}. Also, in other principal components, e.g. PC 4 (Figure 7.6, B, right), the paraffin signals may be a reason for the species-related variation, as can be seen from the presence of the signal at 1460 cm^{-1}. This indicates that the paraffin contribution needs to be minimized before data analysis to obtain classification and identification based on pollen composition alone.

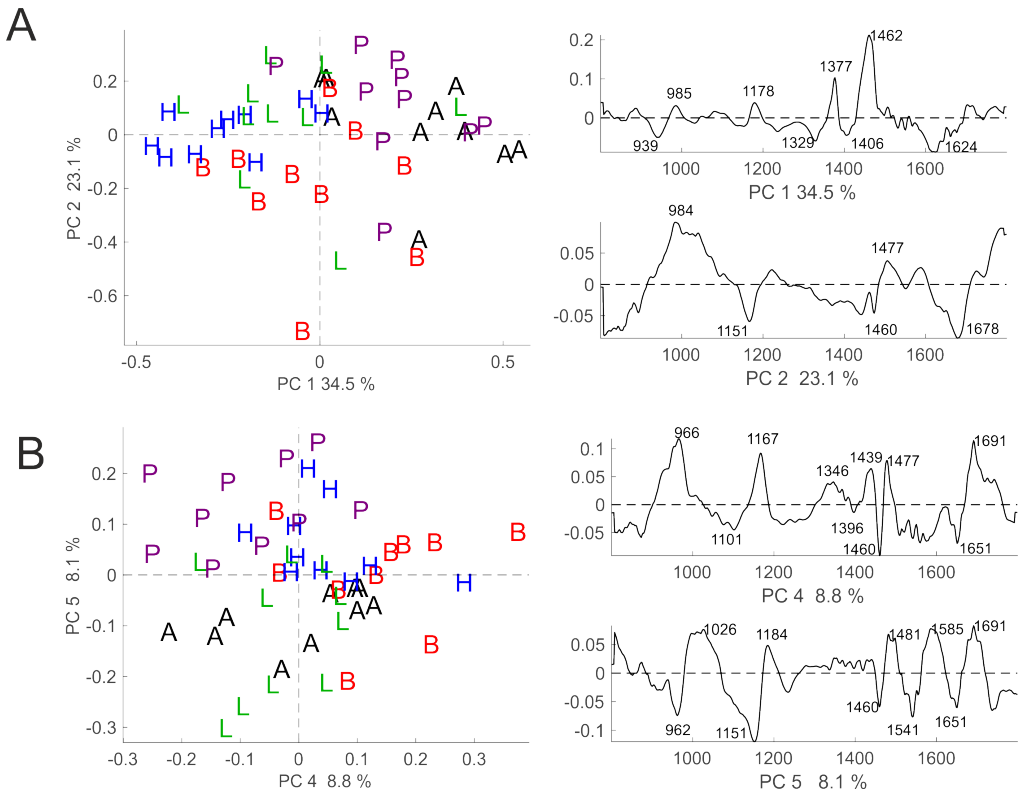

Figure 7.6: Scores plot and loadings from PCA with 50 pollen spectra from the five indicated grass species using the full spectral range from 800 - 1800 cm^{-1}. Each spectrum in the analysis is an average of ~ 20 pollen grain spectra of one individual plant. Spectra were corrected using approach 1 (cf. Figure 7.5, approach 1). **(A)** Scores plot (left) and corresponding loadings (right) of PC 1 and PC 2. **(B)** Scores plot (left) and corresponding loadings (right) of PC 4 and PC 5. Each color and symbol represents the respective grass species. A, *Anthoxanthum odoratum* (black symbols) B, *Bromus inermis* (red symbols) H, *Hordeum bulbosum* (blue symbols) L, *Lolium perenne* (green symbols) P, *Poa alpina* (purple symbols).

7.2.2 Approach 2: Selection of non-affected spectral ranges

Data analysis of concatenated spectra

The spectral region from 1300 to 1500 cm^{-1} is omitted from the spectra of the embedded pollen grains, thus dividing the data in two ranges: 800 - 300 cm^{-1} and 1500 - 1800 cm^{-1}. Before concatenation of the two ranges, each range was baseline corrected separately using AsLS correction.[177,178] After concatenation, EMSC[131] is applied, and Savitzky-Golay smoothing using a windows size of 9 points.[174] For classification by PLS-DA, the individual spectra were used. For analysis by HCA and PCA, averages of the spectra of one respective plant were calculated.

As discussed above, the strong deformation modes of paraffin affect the spectra mostly in the spectral range from 1300 to 1500 cm^{-1} with the two bands at 1377 and 1462 cm^{-1}. Here, this

spectral region was omitted from the data set (compare Figure 7.5, approach 2), similar to the approach in the paraffin-embedding study by Zimmermann *et. al.*[14] Eliminating only the two strongest bands from paraffin, it can be assumed that other spectral features from the paraffin in the regions 800-1300 cm^{-1} and 1500-1800 cm^{-1} are negligibly small compared to the absorption bands of the pollen samples themselves. Following the removal of the 1300-1500 cm^{-1} spectral range, the spectra were normalized by the simple EMSC model.
The assessment by PLS-DA with full-CV (Table 7.2) shows that the overall classification success rate is lower (i.e. 76 %) compared to approach 1, where the contribution by paraffin is not corrected (Table 7.1). Similar to these results, the success rates show high variation for each of the pollen species. They can range from 46 % for *Bromus inermis*, where one-fourth of the actual *Bromus inermis* pollen spectra was misclassified as *Hordeum bulbosum*, to 91 % correct classification of *Anthoxanthum odoratum* and *Poa alpina* spectra.

Table 7.2: Result of PLS-DA classification of 1004 spectra from paraffin-embedded pollen corrected by omitting the spectral range from 1300 to 1500 cm^{-1}, following approach 2 (cf. Figure 7.5). Nine latent variables were used. The results are based on full cross-validation.

affiliation / identified by PLS-DA as	*A. odoratum*	*B. inermis*	*H. bulbasum*	*L. perenne*	*P. alpina*
A. odoratum	181	9	6	16	8
B. inermis	6	91	13	14	6
H. bulbasum	4	52	183	14	0
L. perenne	4	28	1	126	4
P. alpina	4	19	6	26	182
success rates	91 %	46 %	88 %	64 %	91 %
Overall success rate	76 %				

PCA shows that the main variances within this data set are found in the spectral range from 850- 1150 cm^{-1} (Figure 7.7, A, right), which can be assigned mainly to carbohydrates.[4,10] A differentiation between the pollen spectra from *Anthoxanthum odoratum* and *Poa alpina* and between *Hordeum bulbosum* and *Lolium perenne* can be achieved in PC 3 and PC 6, respectively, as shown in the scores plot (Figure 7.7B). The spectral differences in the pollen spectra pre-processed by approach 2 lead to a relatively small drop in classification success rates. This is in agreement with previous work that reports the successful discrimination of paraffin-embedded pollen from other plant species.[14]

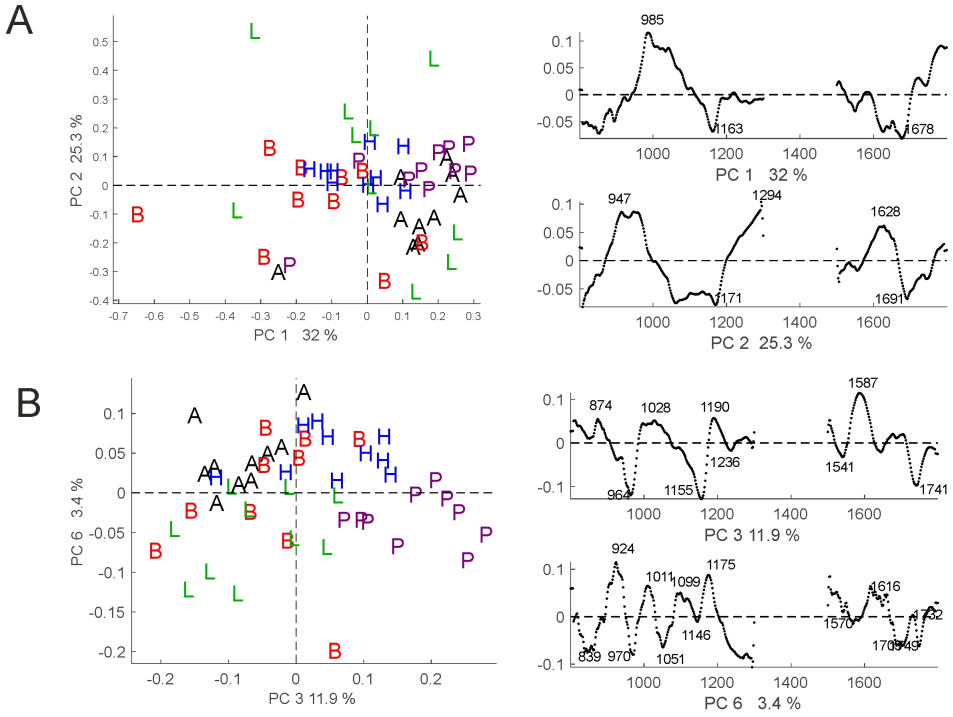

Figure 7.7: Scores plot and loadings from PCA with 50 pollen spectra from the five indicated grass species pre-processed by approach 2 (cf. Figure 7.5, approach 2). Each spectrum in the analysis is an average of the ~ 20 pollen grain spectra of one individual plant. **(A)** Scores plot (left) and corresponding loadings (right) of PC 1 and PC 2. **(B)** Scores plot (left) and corresponding loadings (right) of PC 3 and PC 6. Each color and symbol represents the respective grass species. A, *Anthoxanthum odoratum* (black symbols) B, *Bromus inermis* (red symbols) H, *Hordeum bulbosum* (blue symbols) L, *Lolium perenne* (green symbols) P, *Poa alpina* (purple symbols). To better indicate the region that was excised from the spectra, the loadings are shown as individual data points rather than as line plots.

Combining the spectral ranges 800-1300 cm^{-1} and 1500-1800 cm^{-1} using CPCA

Since the two ranges, 800 to 1300 cm^{-1} and 1500 to 1800 cm^{-1}, can provide different chemical information, namely about polysaccharides in a range from 800 to 1300 cm^{-1} and proteins 1500 to 1800 cm^{-1},[80] the two ranges can be combined using CPCA.[64]

Figure 7.8 shows the resulting scores plots for the global scores (Figure 7.8, A) as well as the block scores (Figure 7.8, B and C) from CPC 1 and CPC 2.

The first and second CPC of the global scores are equally influenced by both ranges, leading to a separation of some pollen spectra with respect to their assigned pollen species (Figure 7.8, A). The first CPC separates the pollen species *Poa alpina*, *Lolium perenne*, and *Hordeum bulbosum* from the spectra of the species *Bromus inermis* and *Anthoxanthum odoratum*. CPC 2 discriminate mainly the spectra from the pollen species *Poa alpina* from the spectra of the other species.

The block scores of the two separate spectral ranges (Figure 7.8, B and C) indicate no discrimination of the score values regarding the different pollen species. Nevertheless, the majority of score values from *Poa alpina* show positive values regarding CPC 1 and CPC 2 and can be differentiated from the score values of the other pollen spectra in both ranges used. In comparison, PCA based on the concatenated spectra of the two spectral ranges leads to a less clear separation of the score values (cf. Figure 7.7). This indicates that a different weighting of the variables is beneficial for separation, due to the relative ratios within bands of the polysaccharide region e.g. the band at 985 cm^{-1} and the protein region with the band at 1680 cm^{-1} (cf. Figure 7.7, loading PC 1).
Since some of the pollen species can be discriminated, the loadings enable characterization of specific patterns for the five different pollen species. Figure 7.9 shows the correlation plot with the loadings of the lower range in black, the values of the upper range in blue, and the centroids of the pollen species in red. For the sake of clarity, just the extrema were presented in the plot. The correlation loadings plot suggests that the separation of pollen species eg. *Poa alpina* and *Bromus inermis* is based on distinct bands in both the lower and the upper range. Pollen spectra of *Poa alpina* have positive global scores for the first and second CPC. Positive correlated bands were at 1186, 1581, 1599, and 1657 cm^{-1} as well as 808, and 1066 cm^{-1} with respect to CPC 2. Most of the bands, namely the bands at 808, 1168, 1580 and 1599 cm^{-1} could be assigned tobuilding blocks of the sporopollenin, such as coumaric acid or ferulic acid.[5,10]

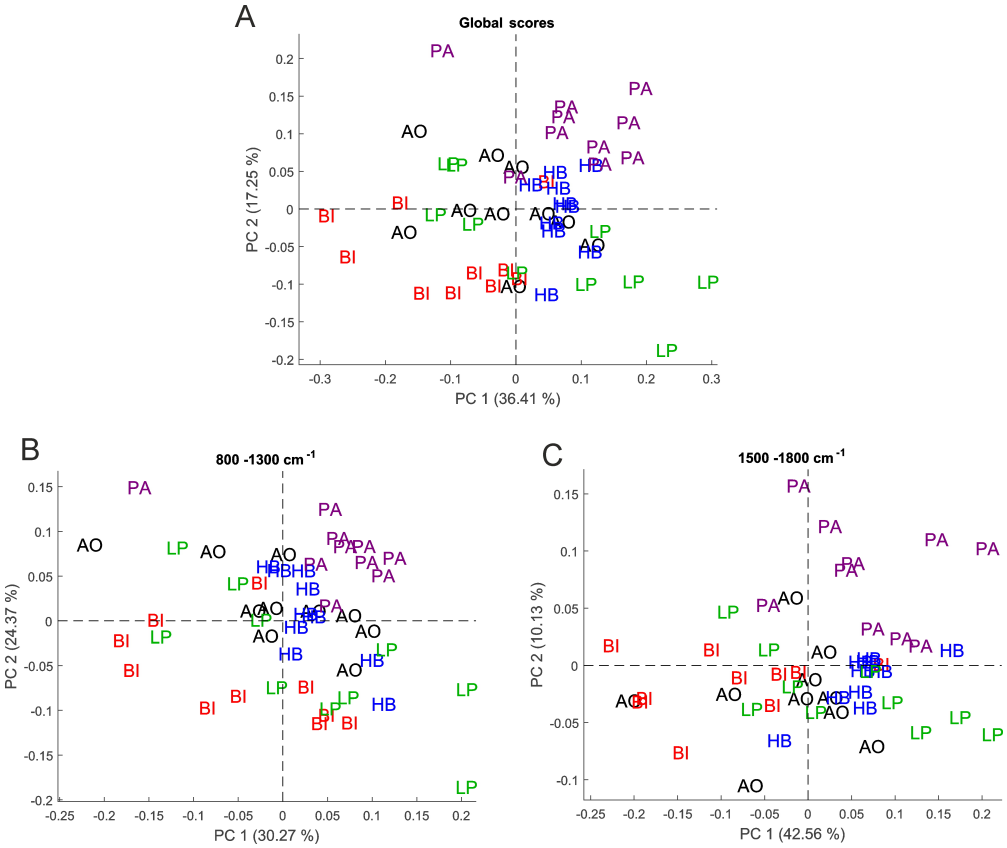

Figure 7.8: (A) Global scores and (**B** and **C**) block scores plots (**(B)** 800 to 1300 cm^{-1} and **(C)** 1500 to 1800 cm^{-1}) of the first and second CPC for averaged FTIR spectra of the grass pollen species *Anthoxanthum odoratum*, AO, black; *Bromus inermis*, BI, red, *Hordeum bulbosum*, HB, blue, *Lolium perenne*, LP, green, and *Poa alpina*, PA, magenta. Spectra were baseline corrected using AsLS correction[177] and normalized using EMSC.[131]

Bromus inermis spectra have negative values regarding both CPC 1 and CPC 2 (Figure 7.8). In agreement with this, the group *Bromus inermis* is highly correlated to the signals at 1157, 1166, 1531, 1689, and 1739 cm^{-1}. These bands are more associated with proteins (1531 and 1689 cm^{-1}) and lipids (1689 and 1739 cm^{-1}).[5,10,80]

Spectra from *Hordeum bulbosum* and *Lolium perenne* have positive scores for the first and negative scores for the CPC 2 (Figure 7.8). The corresponding loadings at 941, 962, 1292 cm^{-1} (Figure 7.9) can be assigned mostly to carbohydrates.[5,10,80]

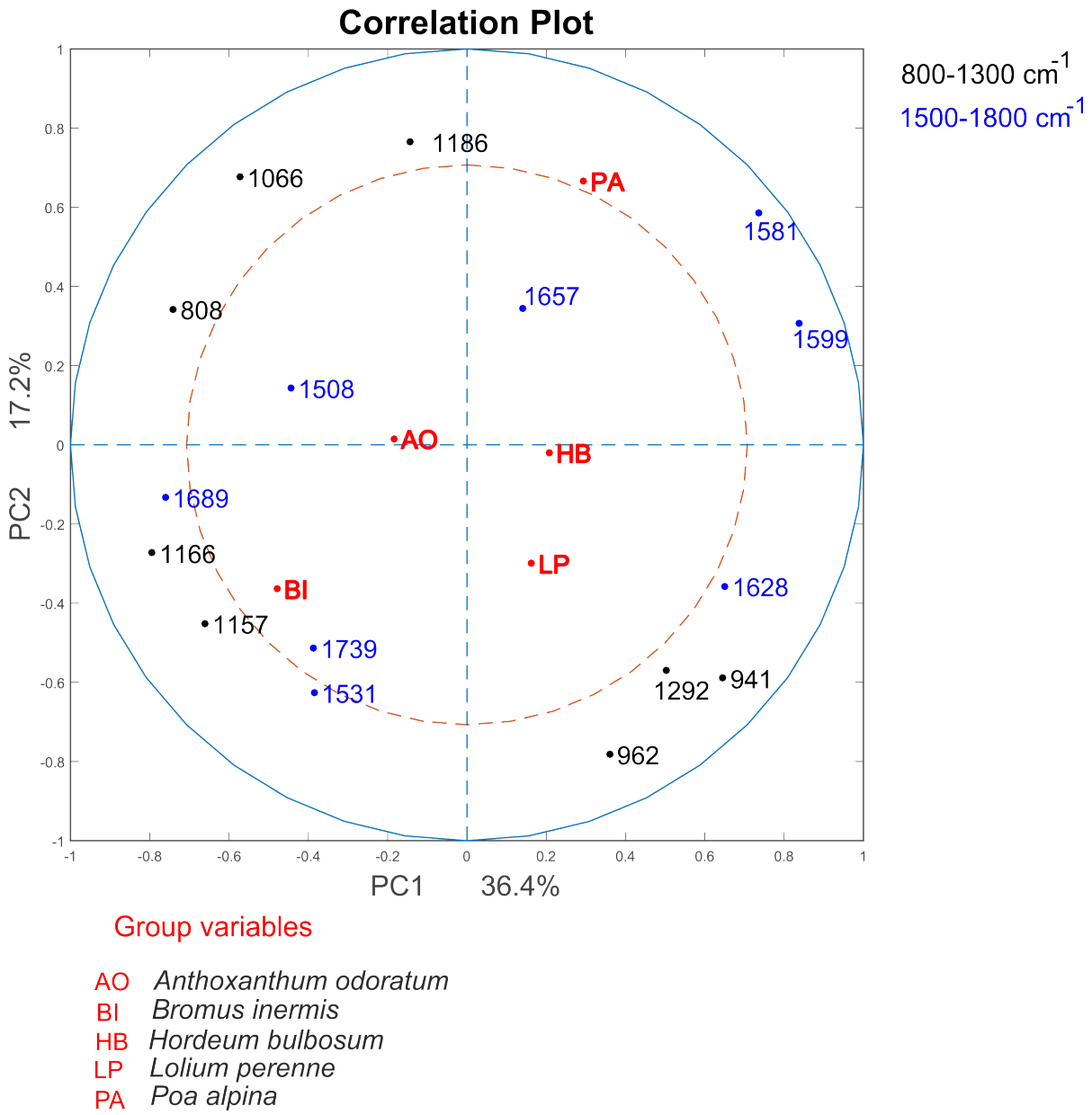

Figure 7.9: Correlation loadings plot with the loadings of the lower range 800 to 1300 cm^{-1} in black, the values of the upper range 1500 to 1800 cm^{-1} in blue, as well the centroids of the pollen species in red. For clarity just the extrema are presented in the plot.

7.2.3 Approach 3: Separation of paraffin and pollen spectra contributions in FTIR spectra of embedded pollen grains

A decomposition of spectral signatures belonging to different chemical constituents within the same spectrum of a complex sample can be achieved by a matrix factorization algorithm, such as NMF. This can result in a more detailed analysis of the spectral features from the different chemical constituents.[219]

The spectra of the embedded pollen grains were baseline-corrected using AsLS correction.

After subsequent vector normalization, NMF was used to split each spectrum into a paraffin and a pollen component to eliminate the paraffin spectral signature. The 1004 pollen spectra and the 190 pure paraffin spectra were decomposed together into six components using the *nnmf*- function in Matlab. All components that contained paraffin signals on visual inspection were separated from those without prominent paraffin signature and left out in the reconstruction of 1004 spectra without paraffin contribution. For classification by PLS-DA, the individual spectra were used. For analysis by HCA and PCA, averages of the spectra of one respective plant were calculated.
Furthermore, matrix factorization algorithms are useful for the exclusion of disruptive contributions from spectra, e.g., for de-noising.[220,221] Therefore, NMF was used in another pre-processing approach (Figure 7.5, approach 3) to split the spectra into pollen spectra and paraffin spectra. In this procedure, the 1004 spectra from each pollen grain and 190 spectra of pure paraffin were decomposed together several times using different numbers of components. The decomposition using six components was chosen as optimal based on visual inspection, which indicated a good separation of the paraffin spectra in components 2 and 6 (Figure 7.10). These two components show the typical paraffin bands at 1377 and 1462 cm^{-1}.

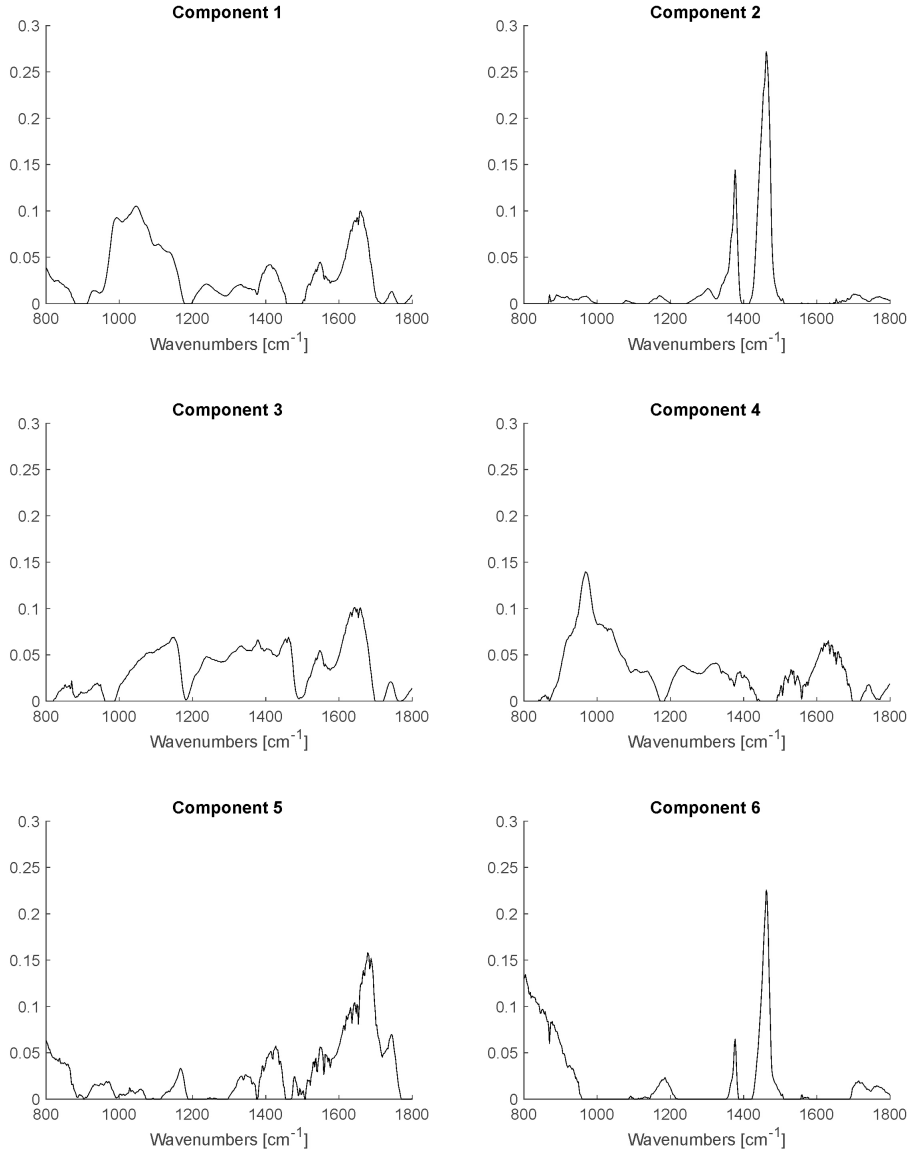

Figure 7.10: Six components of the spectral decomposition by NMF (based on 1004 spectra of paraffin-embedded pollen grains, and 190 pure paraffin spectra, compare Scheme 1, approach 2). Components 2 and 6 show typical contributions by paraffin.

The reconstructed paraffin and pollen spectrum were obtained for each spectrum (each pollen grain), and averages of these two sets of reconstructed spectra for each species are shown in Figure 7.11A and Figure 7.11B, respectively. The reconstructed paraffin spectra (Figure 7.11A) are similar to with a paraffin reference spectrum (Figure 7.11A, top).

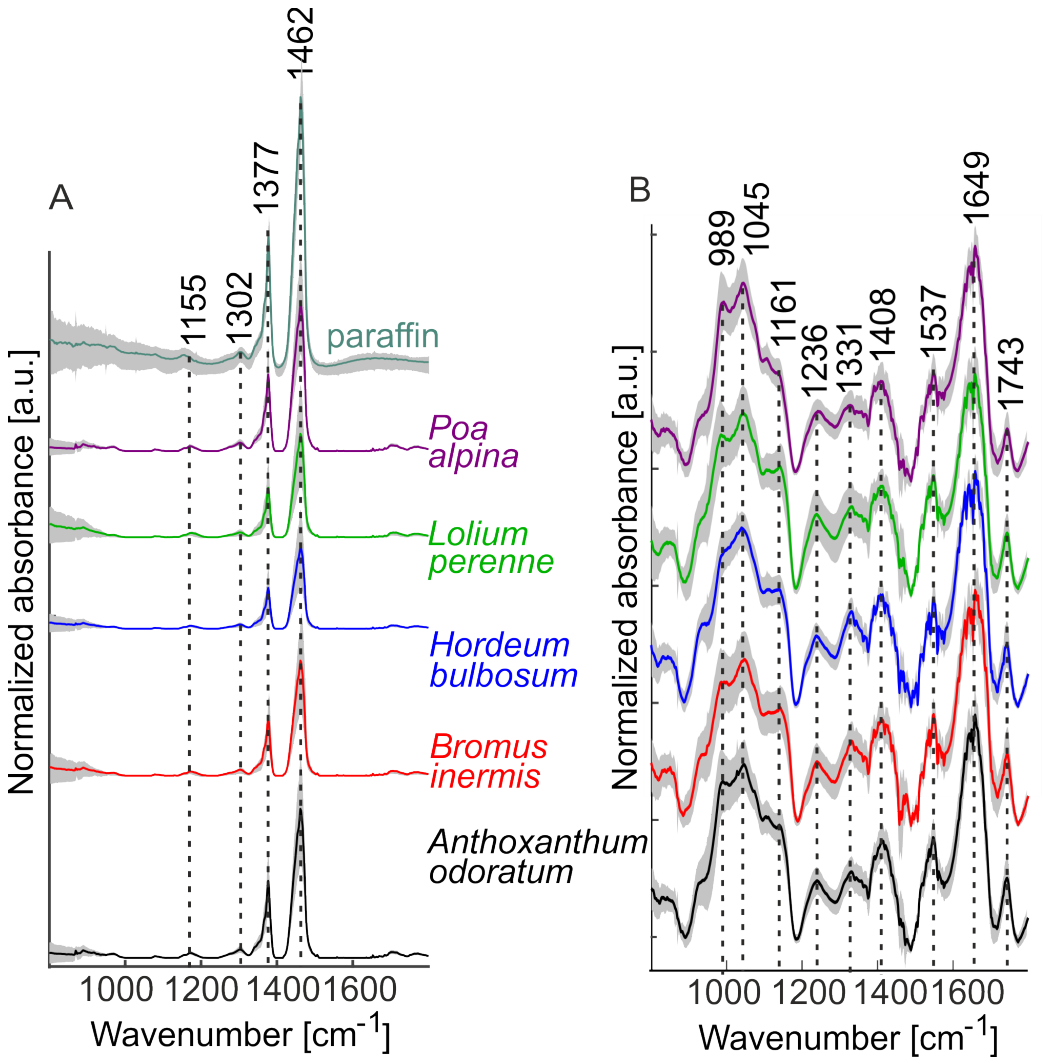

Figure 7.11: (A) Spectra obtained by reconstruction from Component 2 and Component 6 upon NMF with six components (cf. Figure 7.10) for each species, revealing the typical paraffin signature. An average of 190 pure paraffin spectra is shown for comparison (top). **(B)** Reconstructed spectra from NMF components 1, 3, 4, and 5 for each species. All spectra are averages of 200 individual reconstructed spectra (corresponding to 200 pollen grains).

Table 7.3 shows the normalized relative amount of each of the six components. The variation of the relative paraffin contribution (Table 7.3, components 2 and 6) is in good agreement with the visual observation of pollen spectra (Figure 7.4), showing its larger contribution to *Anthoxanthum odoratum* and *Poa alpina* spectra and smaller contribution for the other three species. The averages of the spectra reconstructed from the remaining four components show no characteristic paraffin signals (Figure 7.11B). Compared to the spectra from unembedded single pollen grains on ZnSe slide discussed above (compare Figure 7.4, left), three characteristic bands at 1236, 1331, and 1408 cm^{-1} are more pronounced. They can be assigned to phospholipids, indicated, e.g., by the P=O-stretching vibration at 1236 cm^{-1}; proteins, as illustrated by the C-N stretching mode at 1408 cm^{-1}, and carbohydrates(1331 cm^{-1}) that can be assigned to a ring deformation vibration.[10,80,218]

Table 7.3: Averaged relative spectral contribution of each component after decomposition using NMF (cf. Figure 7.5, approach 3). The spectral contribution is averaged for each pollen species.

spectral contribution / species	Comp 1 [%]	Comp 2 [%]	Comp 3 [%]	Comp 4 [%]	Comp 5 [%]	Comp 6 [%]
A. odoratum	41 ± 12	25 ± 9	12 ± 9	12 ± 11	8 ± 6	2 ± 4
B. inermis	34 ± 13	18 ± 9	15 ± 12	13 ± 9	17 ± 11	3 ± 5
H. bulbasum	36 ± 11	13 ± 7	19 ± 9	17 ± 8	12 ± 8	2 ± 3
L. perenne	34 ± 16	15 ± 10	20 ± 15	18 ± 11	8 ± 7	5 ± 7
P. alpina	42 ± 9	25 ± 9	10 ± 7	13 ± 9	9 ± 6	1 ± 3

The PLS-DA with full-CV classification of the pollen spectra reconstructed by the NMF approach results in a higher success rate (82 %) compared to PLS-DA results of the previous approaches (compare Table 7.4 with Table 7.1 and Table 7.2). The success rates for *Bromus inermis* and *Lolium perenne* are slightly higher (65 % and 71 %, Table 7.4) compared to the classification results of approach 2 (46 % and 64 %, Table 7.2).

Table 7.4: Results of PLS-DA classification of spectra from paraffin-embedded pollen reconstructed from NMF components 1, 3, 4, and 5 (cf. Figure 7.5, approach 3). Nine latent variables were used. The results are based on full cross-validation.

affiliation / identified by PLS-DA as	*A. odoratum*	*B. inermis*	*H. bulbasum*	*L. perenne*	*P. alpina*
A. odoratum	185	2	1	15	12
B. inermis	1	130	9	14	3
H. bulbasum	2	42	192	8	0
L. perenne	6	17	1	140	4
P. alpina	4	8	6	19	181
success rates	93 %	65 %	92 %	71 %	91 %
Overall success rate	82 %				

The corresponding PCA results of the averaged spectra pre-processed by approach 3 (Figure 7.12) shows that the variation within the NMF-decomposed spectra might still be slightly affected by bands from paraffin, as indicated by the variation in the CH_2 deformation at 1460 cm^{-1} that can be assigned to lipids in pollen but could also be present due to residual paraffin contributions[5] (Figure 7.12, right column, loadings of PC 2 and PC 4).

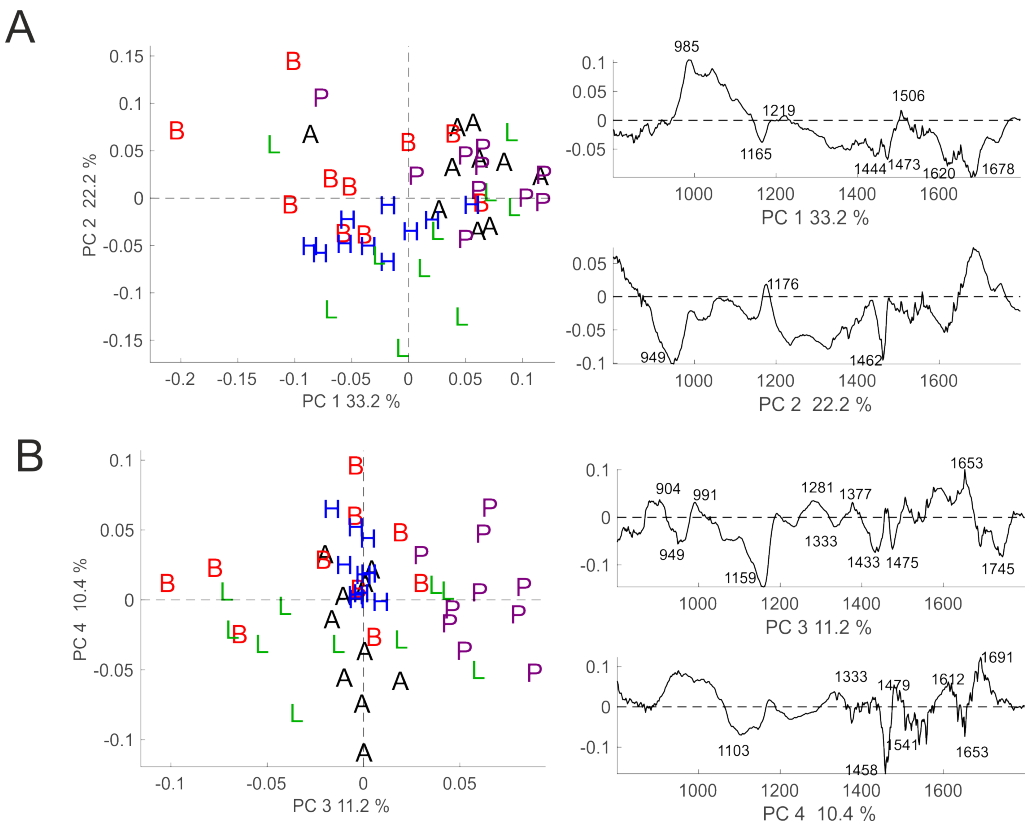

Figure 7.12: Scores plots and loadings from PCA with 50 pollen spectra from five grass species. Each spectrum in the analysis is an average of the 20 pollen grain spectra of one individual plant calculated after reconstruction following a NMF (cf. Figure 7.5, approach 3). **(A)** Scores plot (left) and corresponding loadings (right) of PC 1 and PC 2. **(B)** Scores plot (left) and corresponding loadings (right) of PC 3 and PC 4. Each color and symbol represents the respective grass species. A, *Anthoxanthum odoratum* (black symbols) B, *Bromus inermis* (red symbols) H, *Hordeum bulbosum* (blue symbols) L, *Lolium perenne* (green symbols) P, *Poa alpina* (purple symbols)

7.2.4 Approach 4: correction of the spectra using EMSC with a paraffin constituent spectrum

The AsLS-baseline-corrected spectra of the embedded pollen grains were corrected by the complex EMSC model using a linear and a quadratic component, extended by a representative spectrum of paraffin, as suggested by Kohler *et al.*[222] (Figure 7.5). In contrast to the simple EMSC model used in the pre-processing of the spectra treated by approach 1 and 2 (Figure 7.5), where an average spectrum is used in the model, in the complex EMSC model, two different constituents are estimated in the spectra, specifically the paraffin constituent and the pollen constituent. For the representative spectrum of paraffin for the EMSC model, an average spectrum was calculated from the 190 pure paraffin spectra. For classification by PLS-DA, the individual spectra were used. For analysis by HCA and PCA, averages of the spectra of one respective plant were calculated.

The spectra are normalized so that particularly the bands at 1377 and 1462 cm^{-1} show less variation between the spectra from the five species (Figure 7.13). For the classification, this means that the variation induced by the differences in the paraffin embedding medium can be minimized, and that classification is only based on the spectral contributions by the pollen grains themselves.

The PLS-DA with full-CV classification of the pollen spectra pre-processed by complex EMSC approach results in the highest success rate (83 %) of all the tested approaches (Table 7.5). In particular, the success rate for *Bromus inermis* is higher (63 % in Table 7.5) compared to the classification of the data set corrected using e.g. approach 2 (49 % in Table 7.2). This indicates that the promising classification results obtained in the study by Zimmermann *et. al.* on 11 plant species[14] can be improved even further by optimizing the spectral pre-processing step. In general, approach 3 (the NMF approach, Table 7.4) and approach 4 (the complex EMSC approach, Table 7.5) show relatively similar success rates. For all pre-processing procedures, the success rates vary regarding the different pollen species. The pollen spectra of the species *Anthoxanthum odoratum*, *Hordeum bulbosum*, and *Poa alpina* can be classified more correctly (Table 7.5, >90 %), while identification of the spectra belonging to *Bromus inermis* and *Lolium perenne* is more challenging, with success rates of 63 % and 69 %, respectively.

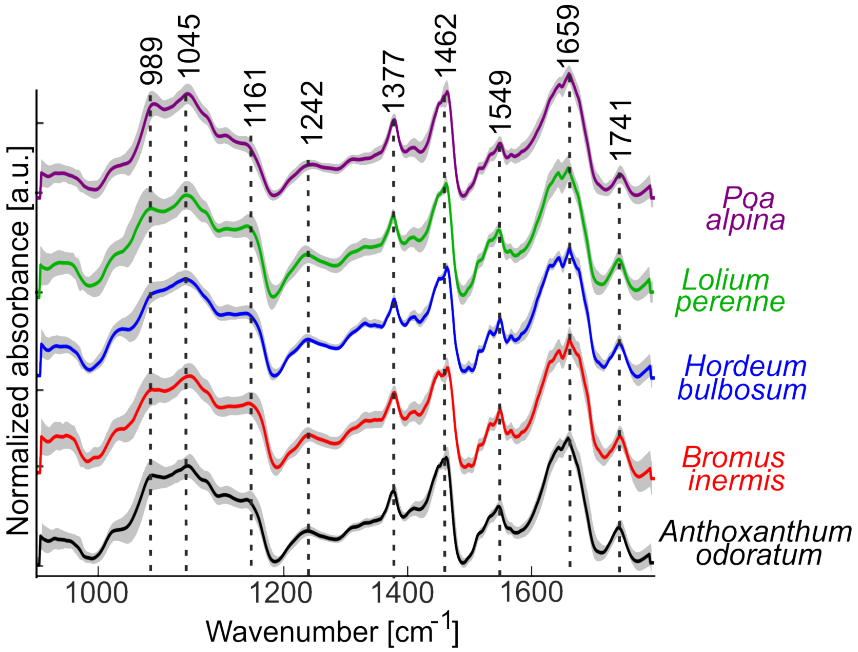

Figure 7.13: FTIR microspectra of paraffin-embedded pollen samples of the five grass species after correction using an EMSC model with paraffin constituent spectrum (cf. Figure 7.5, approach 4). Each spectrum is an average of 200 individual, corrected spectra (corresponding to 200 pollen grains). The standard deviation for each group of spectra is marked in grey.

The success rates for the full cross-validation PLS-DA based classification indicate that the different approaches of minimizing the paraffin contribution to the spectra, namely (i) omitting a part of the spectrum (approach 2), (ii) NMF (approach 3), and (iii) normalization of the paraffin signals by EMSC (approach 4). The approaches lead to similar discrimination of the pollen spectra from the species *Anthoxanthum odoratum, Hordeum bulbosum*, and *Poa alpina*, and less efficient classification of the two species *Bromus inermis* and *Lolium perenne* within the data set. It has been shown before, that spectra of some grass pollen species have more unique spectral features than others so that their discrimination within a data set of similar pollen species is more straightforward.[15,126]

Table 7.5: Results of PLS-DA classification of 1003 spectra from paraffin-embedded pollen corrected using EMSC model with paraffin constituent spectrum (cf. Figure 7.5, approach 4). Eleven latent variables were used. The results are based on full cross-validation.

affiliation / identified by PLS-DA as	*A. odoratum*	*B. inermis*	*H. bulbasum*	*L. perenne*	*P. alpina*
A. odoratum	187	3	1	16	10
B. inermis	1	126	7	14	2
H. bulbasum	2	44	197	7	0
L. perenne	5	19	0	136	4
P. alpina	4	8	4	23	184
success rates	94 %	63 %	94 %	69 %	92 %
Overall success rate	83 %				

The pre-processing of approach 4 has the advantage that no supervision is needed, and automated pattern recognition tools could be developed for fast identification of the spectra. To assess the variances within the spectra, different unsupervised (HCA, PCA) as well as supervised (PLS-DA, Random Forest, ANN) chemometric methods were performed, using the spectral data obtained by approach 4 (Figure 7.5). Subsequently, CPCA of the FTIR data in combination with Raman spectroscopy and MALDI-TOF MS was applied.

7.3 Classification of pollen species using FTIR spectra without paraffin contribution by different chemometric models

7.3.1 Classification by hierarchical cluster analysis and principal component analysis

The hierarchical cluster analysis was carried out with the average spectra of 20 single pollen spectra of each sample, leading to 50 spectra in total, using Euclidean distances and Ward's algorithm. The resulting dendrogram is shown in Figure 7.14. Most of the spectra of *Poa alpina* (Figure 7.14, purple branches), *Anthoxanthum odoratum* (Figure 7.14, black branches), and *Hordeum bulbosum* (Figure 7.14, blue branches) are clustered almost exclusively in their respective groups. This is in good agreement with the PLS-DA identification discussed in Table 7.5 and indicates low variances within the spectra of the respective species. The high similarity of the majority of the spectra from *Bromus inermis* (Figure 7.14, red branches) to those of *Hordeum bulbosum* (Figure 7.14, blue branches) agrees with the high number of *Bromus inermis* spectra that are misclassified in the PLS-DA as *Hordeum bulbosum* spectra (cf. Table 7.5). Therefore, it can be concluded that the composition of the pollen grains of the two species is very similar, in agreement with the close relationship of the tribes Hordeeae (Triticeae) and Bromeae within the Pooideae subfamily.[223,224]

The cluster in the dendrogram that comprises all except one spectrum from *Poa alpina* (Figure 7.14, purple branches) is very similar to a group of spectra that contains average pollen spectra from *Anthoxanthum odoratum* and *Lolium perenne* plants (Figure 7.14, black branches and green branches, respectively), also in agreement with the misclassification by PLS-DA of several individual spectra from these species (Table 7.5). Moreover, it can be concluded that the chemical composition of their pollen is similar compared to those from the other species, in agreement with the fact that all of them belong to the Poeae/Aveneae tribe complex.[224]

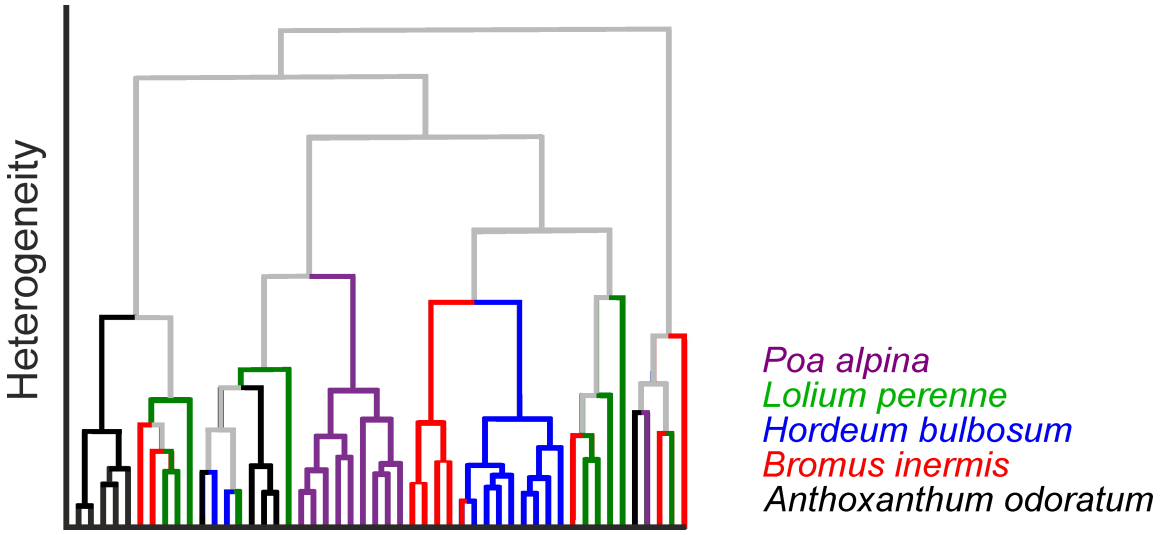

Figure 7.14: Dendrogram obtained after hierarchical cluster analysis (HCA) with 50 pollen spectra from the five indicated grass species, using the full spectral range from 800-1800 cm^{-1}. Each spectrum in the analysis is an average of the $\tilde{2}0$ pollen grain spectra of one individual plant). HCA was executed using Euclidean distances and Ward's algorithm.[58] The colored branches correspond to the font color with which the respective pollen species is listed.

In a PCA, the differences between the spectra of the five pollen species can be identified based on the loadings spectra. Figure 7.15A shows the scores plot and corresponding loadings of the first and second principal components of a PCA with the averaged pollen spectra of the 50 plants. The first and second PC explain 54 % of the total variance in the data set. As visible in the scores plot in Figure 7.15A, the mostly positive score values regarding the first PC of spectra from *Poa alpina* and *Anthoxanthum odoratum*, as well as most spectra from *Lolium perenne* confirms the high similarity of the pollen composition in these two species. The spectra from *Bromus inermis* and *Hordeum bulbosum* show mostly negative score values regarding the first PC (Figure 7.15A).

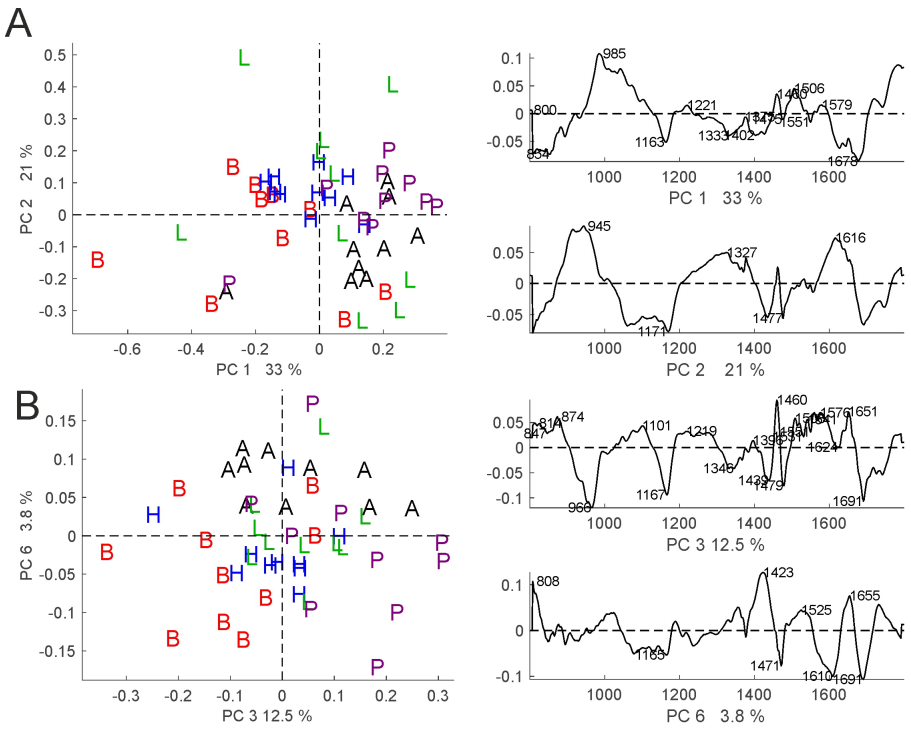

Figure 7.15: Principal component analysis (PCA) of 50 pollen spectra from the five indicated grass species. **(A)** Scores plot and corresponding loadings of PC 1 and PC 2. **(B)** Scores plot and corresponding loadings of PC 3 and PC 6. Each color representing the respective pollen species. Abbreviations: A, *Anthoxanthum odoratum* (black symbols), B, *Bromus inermis* (red symbols), H, *Hordeum bulbosum* (blue symbols), L, *Lolium perenne* (green symbols), P, *Poa alpina* (purple symbols).

According to the loadings in Figure 7.15A, the most pronounced differences between the spectra from the *Bromus inermis*/ *Hordeum bulbosum* group and those from the two species *Poa alpina* and *Anthoxanthum odoratum* are in the broad bands around 945 cm^{-1} (Figure 7.15A, second PC) and 1678 cm^{-1} (Figure 7.15A, first PC) that could be assigned to molecular vibrations of carbohydrates and proteins, respectively.[4,80,218] This would lead to the conclusion that pollen from *Bromus inermis*/ *Hordeum bulbosum* can be discriminated from *Poa alpina* and *Anthoxanthum odoratum* based on different carbohydrate and protein composition. The scores plot in Figure 7.15B shows that separation of *Poa alpina* and the *Bromus inermis* / *Hordeum bulbosum* group from the other species is also possible based on the variance in the third PC. According to the corresponding loading spectra in Figure 7.15B, the discrimination is achieved by signals that can be assigned to carbohydrates at 966 cm^{-1}, to sporopollenin at 1167 cm^{-1}, and 1610 cm^{-1}, tentatively assigned to lipids at 1423 cm^{-1} and 1460 cm^{-1}, and proteins at 1651 cm^{-1} and 1691 cm^{-1}.[5,10]

7.3.2 Pattern recognition for classification of grass pollen spectra from independent populations

A robust, reliable discrimination method should include the possibility to identify pollen spectra that come from different plant populations rather than from the same set of plant populations. Therefore, in this experiment, the plants in each of the five species originated from two different populations. A PLS-DA model was constructed using spectra from just one population per species, amounting to 502 spectra. The success rates were obtained by using the respective other population from each species as an independent test set, comprising other 502 spectra. The results for the identification of the unknown populations by PLS-DA are shown in Table 7.6. Comparing the success rates with the results obtained full-CV approach shown above (Table 7.5), the success rates are only slightly lower for the species *Anthoxanthum odoratum*, *Hordeum bulbosum*, and *Poa alpina* when spectra come from an unknown population. Nevertheless, the success rates for those species are very low, where success rates are already low in the full-CV identification, that is, in *Bromus inermis* and *Lolium perenne* (compare Table 7.6 with Table 7.5), with the success rate for classification of the former drops from 63 % to 26 %.
Identification using a random forest algorithm results in similar success rates as the PLS-DA model in the case of *Bromus inermis* and *Lolium perenne*, but lower numbers of correct identification than PLS-DA (Table 7.6) for the other three species (Table 7.7). Changing the number of trees in the RF from 300, which was determined to be optimum to higher numbers, results in similar success rates.

Table 7.6: Results of the PLS-DA classification of the FTIR spectra. Training set was trained with 9 latent variables. Training of the classification models was based on spectra from only one population for each grass species, while the independent validations were conducted using the other respective population for each species.

affiliation / identified by PLS-DA as	*A. odoratum*	*B. inermis*	*H. bulbasum*	*L. perenne*	*P. alpina*
A. odoratum	82	4	0	12	1
B. inermis	0	24	6	4	4
H. bulbasum	8	27	94	7	2
L. perenne	4	36	1	46	4
P. alpina	6	7	3	31	87
success rates	82 %	26 %	90 %	46 %	89 %
Overall success rate	67 %				

Table 7.7: Results of the Random forest classification of the FTIR spectra. Training set was trained with 300 bags. Training of the classification models was based on spectra from only one population for each grass species, while the independent validations were conducted using the other respective population for each species.

affiliation / identified by RF as	*A. odoratum*	*B. inermis*	*H. bulbasum*	*L. perenne*	*P. alpina*
A. odoratum	75	31	2	23	6
B. inermis	1	26	10	4	15
H. bulbasum	6	4	87	1	3
L. perenne	10	34	2	47	2
P. alpina	8	5	3	25	72
success rates	75 %	26 %	84 %	47 %	73 %
Overall success rate	61 %				

A feed-forward artificial neural network (ANN) was trained with the same set of 502 spectra, divided into a training, validation, and internal test set. The net is tested with 502 spectra from the other respective populations. The success rates were very similar, with a higher number of correct species assignments in *Lolium perenne* and similar misclassification, e.g., assignment of *Lolium perenne* as *Poa alpina* (Table 7.8). The slightly diminished success rate for the identification of *Hordeum bulbosum* compared to the PLS-DA classifier is more accurate, considering a 66 % correct identification of the spectra from *Lolium perenne* pollen that is an improvement compared to the PLS-DA model (compare Table 7.6 and Table 7.8). Consistent with the results of HCA (Figure 7.14) and PCA (Figure 7.15), almost all incorrectly assigned spectra of *Hordeum bulbosum* are labeled with *Bromus inermis* as output class, also in agreement with the close phylogenetic relationship of the two species mentioned above.[223,224]

Table 7.8: Results of the ANN classification of the FTIR spectra. Training set was trained with 519 input variable and 50 hidden units. Training of the classification models was based on spectra from only one population for each grass species, while the independent validations were conducted using the other respective population for each species..

affiliation / identified by ANN as	*A. odoratum*	*B. inermis*	*H. bulbasum*	*L. perenne*	*P. alpina*
A. odoratum	80	1	0	4	1
B. inermis	0	30	20	4	5
H. bulbasum	10	22	81	2	1
L. perenne	6	39	0	66	2
P. alpina	4	8	3	24	89
success rates	80 %	30 %	77 %	66 %	91 %
Overall success rate	83 %				

The strong decrease in classification success in *Bromus inermis* when an unknown population must be identified and the high success rates for other species are in agreement with the different intra-species variance that was observed between populations of other Poaceae species.[15] Especially in *Anthoxanthum odoratum* and also *Poa alpina* that show the highest success rates, the ability to distinguish spectra from different populations of the same species was challenging based on FTIR spectra[15] but could be achieved using other chemical information of the pollen samples in a multiblock approach.
The fact, that identification is based on spectra from individual pollen grains rather than averages from one plant adds another source of variation, is also discussed in chapter 5 when different spectroscopic methods that probe either bulk samples or individual pollen grains and their potential for pollen identification are compared. Nevertheless, the possibility to study pollen spectra in mixtures could in the future open possibilities for the FTIR-imaging based identification of mixed grass pollen samples, similar to existing high-throughput and mapping approaches.[107,148]

7.4 Combination of FTIR spectra with Raman spectra and MALDI mass spectra for the classification and characterization of pollen from different grass species

As discussed in chapter 5 the combination of pollen spectra obtained by complementary methods leads to significantly better discrimination of different pollen populations. In Chapter 4 the ability of MALDI MS spectra to discriminate between different pollen species was demonstrated. Starting from the set of species that was analyzed based on FTIR spectral information from paraffin-embedded single pollen grains, also the ablility of Raman spectroscopy and MALDI-MS to help in this specific classification problem will be assessed. Here, the same sample set that was used in the discussion of the previous sections of this chapter (Figure 7.1, first level) was measured using MALDI TOF MS. Unfortunately, in a MALDI-MS experiment, a higher amount of pollen grains is needed[1] than in the microspectroscopic methods, therefore, only 49 spectra of sufficient quality out of 50 samples were obtained.
The 50 samples were measured and analyzed using Raman spectroscopy by Simon Schröder at a part of a research internship (unpublished data). The 49 corresponding averaged spectra were utilized in a comprehensive CPCA using three blocks, FTIR spectra, Raman spectra, and MALDI mass spectra.

7.4.1 Separate analysis of the MALDI-MS data

Before the combination of the FTIR microspectra with the MALDI-MS data, the MALDI-MS data were assessed separately regarding their application to discriminate between the five grass species in this sample set. A PLS-DA model using full-CV and 4 latent variables was applied to assess the discrimination based on the spectrometric fingerprint obtained by MALDI TOF MS. Table 7.9 shows the classification results of the spectra from the five plant species. As expected from the previous results, the high success rates listed in Table 7.9 for the sample set Pollen Norway IIa confirms the high species-specificity of MALDI-TOF MS. From 49 spectra only 2 spectra were misclassified, namely one spectrum of *Lolium perenne* and one spectrum of *Poa alpina*, which yields to an overall success rate of 94 %. Both were misclassified as *Hordeum bulbosum*.

Table 7.9: Results of PLS-DA classification of 49 mass spectra from grass pollen. 4 latent variables were used. The results are based on full cross-validation.

affiliation / identified by PLS-DA as	*A. odoratum*	*B. inermis*	*H. bulbasum*	*L. perenne*	*P. alpina*
A. odoratum	10	0	0	0	0
B. inermis	0	10	0	0	0
H. bulbasum	0	0	10	1	1
L. perenne	0	0	0	9	0
P. alpina	0	0	0	0	8
Suceess rates	100 %	100 %	100 %	90 %	89 %
Overall success rate	96 %				

Furthermore, PCA was applied to the data set of 49 mass spectra. Figure 7.16 displays the scores plots of the first and second PC as well as the scores plot of the third and fourth PC and indicates that the variances caused by the species discrimination can be explained by using PC1 to PC4. The first PC, which explains 19.8 % of the total variance described the separation of spectra from *Anthoxanthum odoratum* (Figure 7.16A, black symbols) from the rest of the spectra. The loadings for PC 1 show high signals in the peaks of *m/z* 2256, 5228, and 5742, which are dominant in the averaged spectra as well (cf. Figure 4.2). All spectra of *Poa alpina* (Figure 7.16A, purple symbols) have positive values regarding PC 2. The explained variance of PC2 with 15.3 % is slightly lower than the explained variance of PC 1, which can lead to the conclusion, that the spectra of *Poa alpina* are also highly species-specific. The loadings of PC2 point out two signals at *m/z* 2390 and 3428 that describes the species-specific pattern of *Poa alpina*.

PC 3 causes the separation between spectra from *Bromus inermis* (Figure 7.16B, red symbols) and the rest of the spectra and explains 12.7 % of the total variance. The loadings indicate that one dominant peak at m/z 1222 causes the separation. PC4 enables discrimination of

pollen mass spectra from *Lolium perenne* from spectra of the other species (Figure 7.16B, green symbols). Also here, two dominant peaks at *m/z* 2402 and 4970 can explain the discrimination from spectra of *Lolium perenne* (Figure 7.16, purple symbols) and the rest of the spectra.

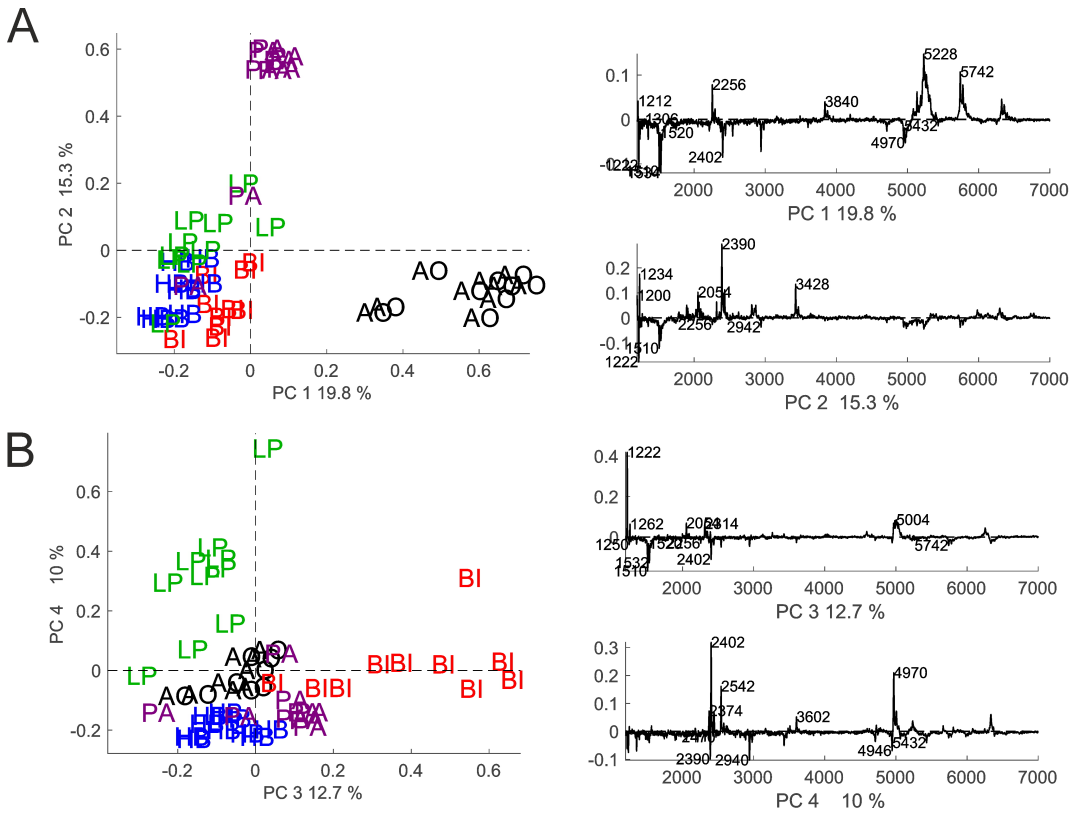

Figure 7.16: PCA of 49 pollen spectra from the five indicated grass species. **(A)** Scores plot and corresponding loadings of PC 1 and PC 2. **(B)** Scores plot and corresponding loadings of PC 3 and PC 4. Each color represents the respective pollen species. Abbreviations: AO, *Anthoxanthum odoratum* (black symbols), BI, *Bromus inermis* (red symbols), HB, *Hordeum bulbosum* (blue symbols), LP, *Lolium perenne* (green symbols), PA, *Poa alpina* (purple symbols).

MALDI mass spectra of *Hordeum bulbosum* (Figure 7.16, blue symbols) have mostly negative values for the first to fourth PC. Each component from PC1 to PC4 separates one pollen species from the other pollen species. Higher PCs were investigated, where a separation also for *Hordeum bulbosum* spectra can be determined (PC 5, 5% explained variances), but with a higher distribution over the principal component. As will also be discussed in Chapter 8 of this thesis, the knowledge about species-specific peaks within the pollen spectra offers the application of multiplexing[225] and MALDI mass imaging[56] as long as the sample extracts do not overlap.[55] Here, MALDI-MS is used to complement the pre-processed FTIR microspectra and Raman microspectra in a CPCA. The pre-processing applied to the FTIR microspectra was performed according to approach 4, as discussed in detail in section 7.2 (see Figure 7.5).

7.4.2 Results of CPCA combining FTIR microspectra with Raman and MALDI-MS data

Figure 7.17 shows the global scores and blocks scores of the first and second CPCs. Regarding the global pattern, data from the three different pollen species *Bromus inermis*, *Anthoxanthum odoratum*, and *Poa alpina* can be distinguished from each other concerning CPC1. This first component explains 22 % of the variances and is mostly influenced by the Raman block. The data from *Poa alpina* is separated by CPC1 in the Raman and MALDI block, in addition to the global scores. Furthermore, *Poa alpina* data in the block scores of the FTIR block result in mostly positive score values, that separates the data from Poa alpina from the score values from *Bromus inermis* spectra.
It should be pointed out, that the appearance of the FTIR block scores plot differs from the single PCA in Figure 7.15, while the MALDI block scores plot show high similarity to the single PCA (Figure 7.16).

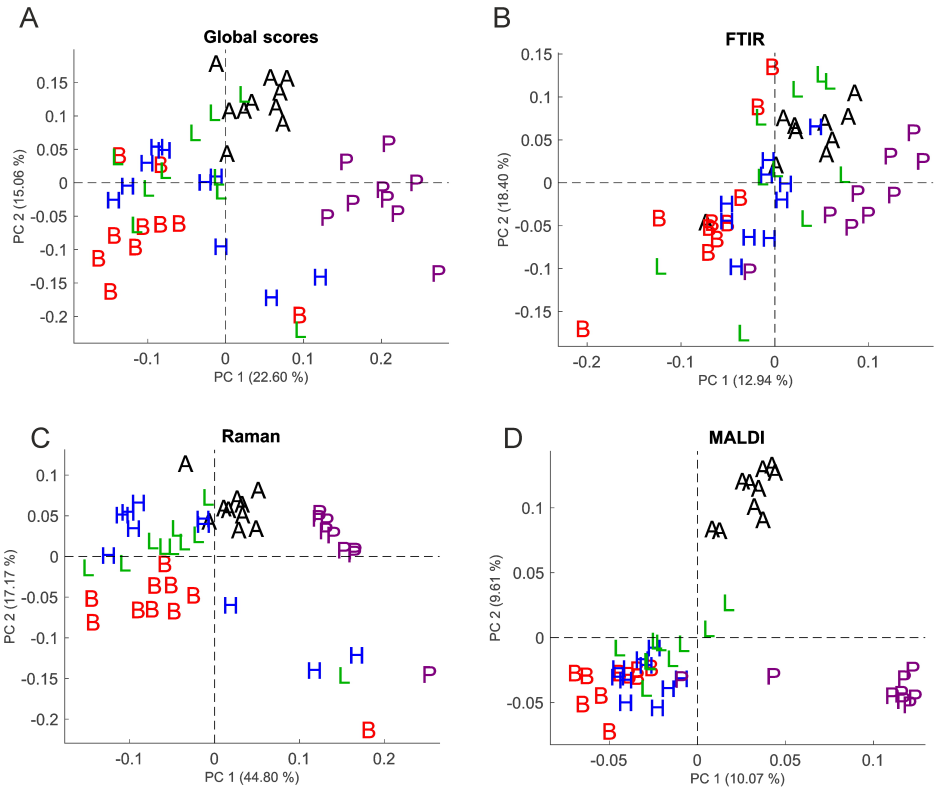

Figure 7.17: (A) Global scores plot and **(B-D)** blocks score plots (**(B)** FTIR , **(C)** Raman, and **(D)** MALDI) of the first and second CPC for averaged pollen FTIR spectra of *Anthoxanthum odoratum*, A, black; *Bromus inermis*, B, red, *Hordeum bulbosum*, H, blue, *Lolium perenne*, L, green, and *Poa alpina* , P, magenta

The loadings are represented in a correlation loadings plot. In Figure 7.18 the group variables represent the five different pollen species and the extrema of the loadings from the methods are shown. The global score values of *Anthoxanthum odoratum* are positively correlated to

the MALDI bands at *m/z* 2256, 5230 and 1212. The peaks at *m/z* 2256 and *m/z* 5230 were already revealed as species-specific in the discussion above (compare Figure 7.16, A). The score values of *Anthoxanthum odoratum* are positive correlated to the FTIR bands at 989, 1458, and 1508 cm^{-1}, which can be assigned to carbohydrates (989 cm^{-1}) and proteins(1458 and 1508 cm^{-1}).[10]

The score values of the *Poa alpina* data are highly correlated with MALDI peaks at *m/z* 1234, 2390 and 3428. Like in the case of *Anthoxanthum odoratum* these peaks are also highly pronounced in the loadings of Figure 7.16, A. FITR bands at 804, 1279, 1373, and 1579 cm^{-1}, as well as the Raman bands at 420, 503, 648, and 968 cm^{-1} show also a correlation to the score values of *Poa alpina*. The bands can be tentatively assigned to the molecular vibrations of proteins.[80]

The global score values of *Bromus inermis* show a highly positive correlation to the MALDI species-specific peak at *m/z* 4984. The dominant signal of the peak at *m/z* 2402 in the loadings of the fourth PC (Figure 7.16, B) is not visible in the correlation loadings plot and therefore not highly correlated after the weighting of CPCA using FTIR and Raman. FTIR bands at 852, 1333, and 1689 cm^{-1}, as well as the Raman bands at 468 and 935 cm^{-1}, are correlated to the scores of *Bromus inermis*. These bands can be assigned to carbohydrates, such as starch.[16,203]

The centroids of the score values of *Lolium perenne* and *Hordeum bulbosum* are located closer to the origin of the correlation loadings plot, which indicates, that their species-specific variances are less explained by the plot. This is caused by a group of outliers, namely two score values of *Hordeum bulbosum*, one of *Lolium perenne*, one of *Bromus inermis*, and one of *Poa alpina*, mainly in the Raman block (Figure 7.17. The global score values of the two species *Lolium perenne* and *Hordeum bulbosum* can hardly be discriminated from each other and they are correlated with MALDI peaks at *m/z* 1222, 1510, and 2940, and with the FTIR bands at 1161, 1427, and 1550 cm^{-1} that are possible constituents of sporopollenin.[5,10]

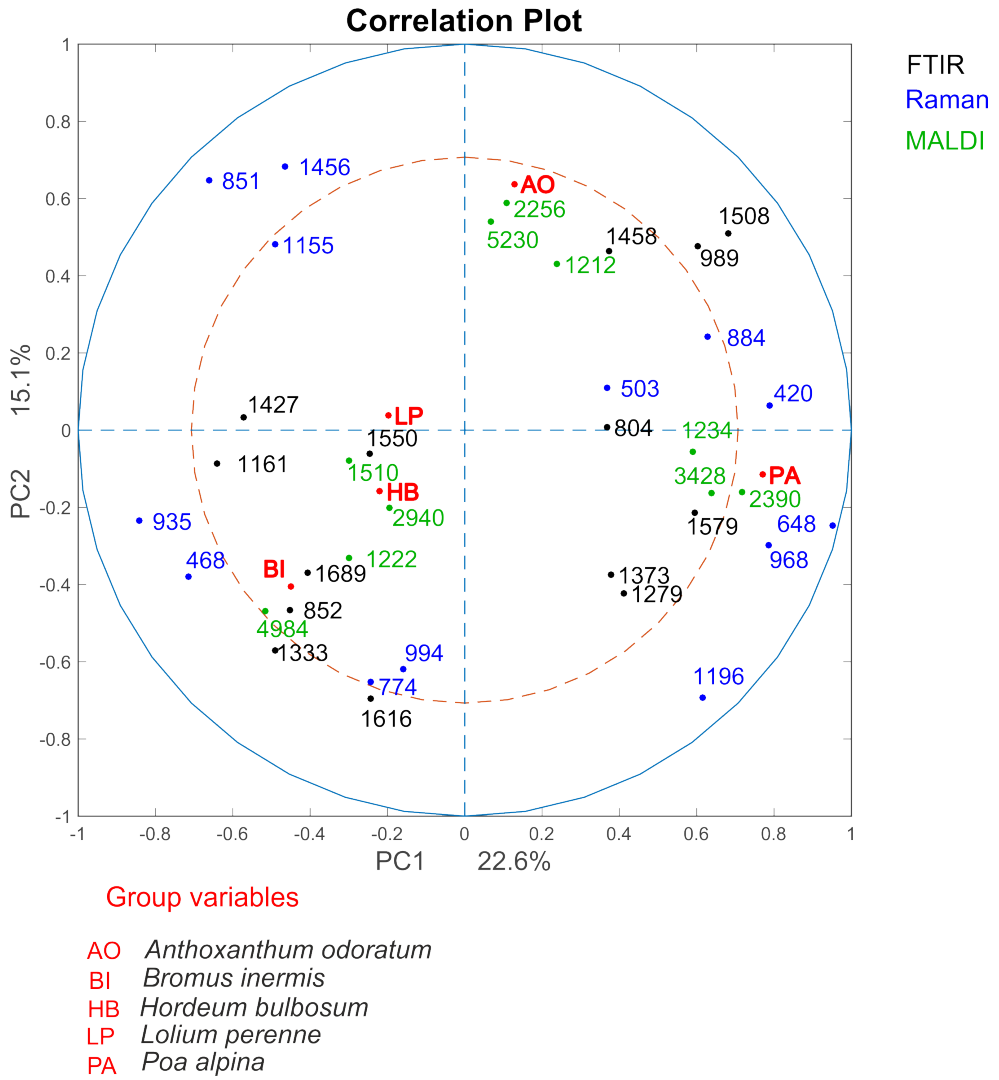

Figure 7.18: Correlation loadings plot with the loadings of the FTIR black, Raman, blue and MALDI, green, as well the centroids of the five pollen species in red. For the sake of clarity just the extrema were presented in the plot.

7.5 Reproducibility of FTIR data of individual pollen grains

In the previous sections of this chapter, the evaluation of classification models was discussed using five different pollen species measured within a short time (FTIR, 10 days, MALDI, 1 day). One criterion for the development of a suitable classification approach using chemometric methods is the reproducibility of the measurements. Data might be obtained at a later time, and new samples need to be added into an existing data set.

Several studies were conducted to analyze and understand the low reproducibility in FTIR experiments.[226] This section will address these challenges regarding reproducibility of single pollen grain FTIR microspectra and compare them to those of MALDI MS experiments on the extracts of the same sample set. First, the model is validated by the technical replicates

of the identical sample set, measured at a later time (30 samples from Pollen Norway IIa) and second, a sample set comprising additional grass species species (Pollen Norway IIb) measured at a different time is added to the existing data set for both FTIR and MALDI.
Figure 7.19 gives an overview of the taxonomical relation in the complete sample set Pollen Norway II (a (Figure 7.19, blue) +b (Figure 7.19, green)).
The taxonomical variation with Pollen Norway IIa consists of three pollen species from the tribe Poaeae, namely *Lolium perenne, Poa alpina*, and *Anthoxanthum odoratum*. Furthermore, the pollen species *Hordeum bulbosum* and *Bromus inermis* belong to the tribe Bromeae and Triticeae respectively are both tribes of the same supertribe Triticoadae.
The sample set Pollen Norway IIb adds two more pollen species of the tribe Triticeae, namely *Hystrix patula* and *Hordeum vulgare*, as well as the two pollen species *Piptatherum millaceum* and *Piptochaetium avenaceum* from the tribe Stipeae.[27]

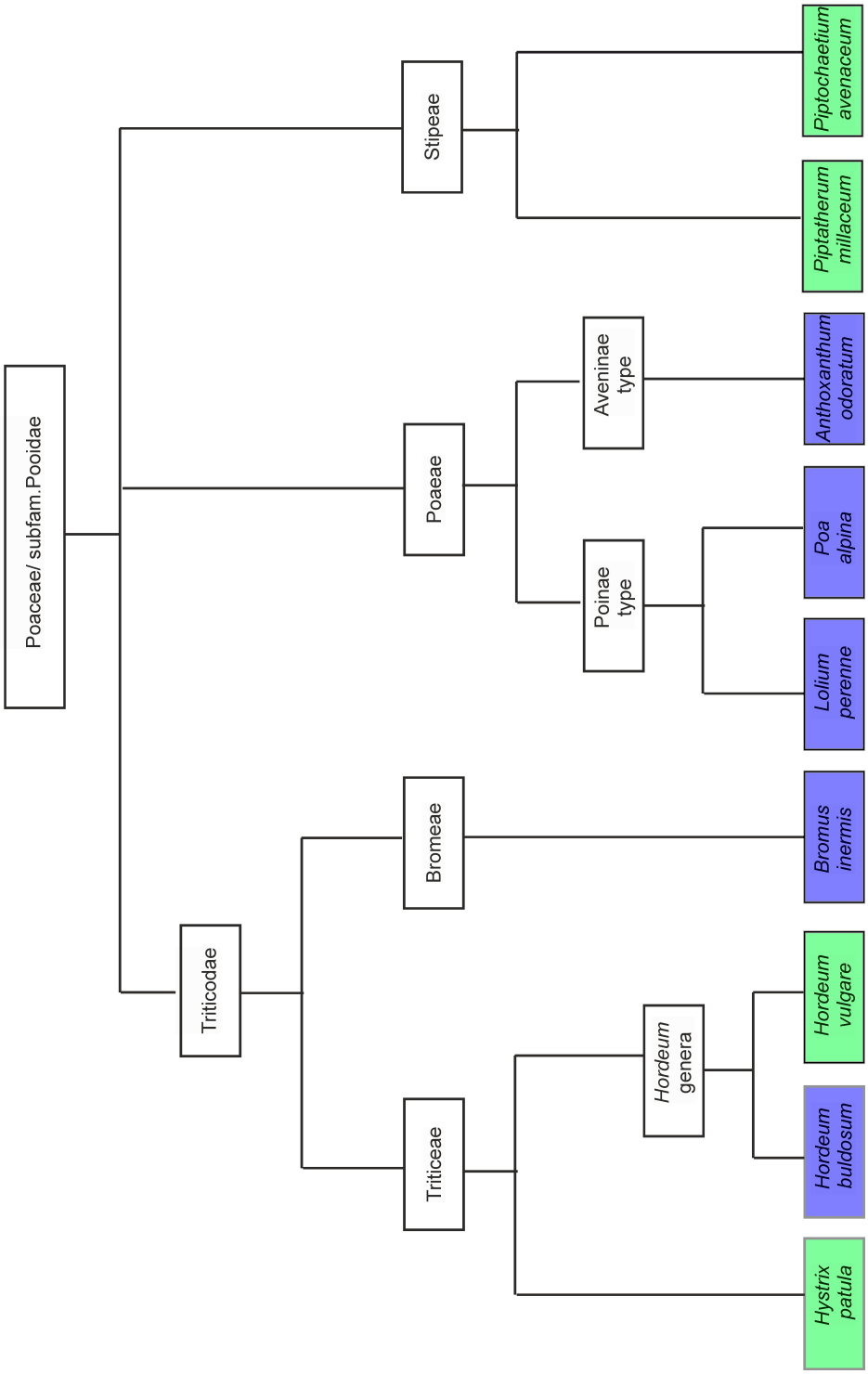

Figure 7.19: Taxonomical relation of the nine pollen species of the sample set Pollen Norway IIa (blue) and Pollen Norway IIb (green) based on the review by Soreng *et al.*[27] With the first level ist giving the family/subfamily, followed by the supertribes as the second level and the tribes, third level Poacea pollen are described in two cluster Poidae type and Aveninae type.

7.5.1 Repeated FTIR experiment with the same sample set

30 samples from the sample set Pollen Norway IIa, more precisely six samples for each pollen species were measured using FTIR and the same embedding procedure described above. To assess the reproducibility, all 600 spectra of the additional measurements were used as an independent test set for PLS-DA validation. For this classification, the concatenation approach discussed above (Figure 7.5, approach 2) was used for pre-processing the data. This approach shows similar classification results compared to approach 3 (NMF, Table 7.4) and approach 4 (complex EMSC, Table 7.5), but do not require additional paraffin spectra or complex modeling.
The classification results are presented in Table 7.10, the overall success rate is very low with 40 %. For *Anthoxanthum odoratum* more than 60 % of the spectra are correctly classified and more than 80% of the spectra are classified correctly in the case of *Poa alpina*. It should be pointed out, that most of the misclassified spectra of *Anthoxanthum odoratum* are classified as *Poa alpina* and *vice versa*, , in agreement with the observation in the first test (complex EMSC, Table 7.5).
The success rates for the three other pollen species *Bromus inermis*, *Hordeum bulbosum*, and *Lolium perenne* are low. Most of the spectra from the independent test set were incorrectly classified as *Anthoxanthum odoratum*.

PCA was applied to the averaged spectra of the first 30 samples (A) and the new 30 samples (B). The scores plot in Figure 7.20, left indicates an influence of the different measuring times within the first and second PC that explain together almost 60 % of the total variance in the data. The loadings in Figure 7.20, right, show that the variance is mostly caused by FTIR bands at 916, 980, 1049, and 1163 cm^{-1}. These bands can be assigned to molecular vibrations of carbohydrates.[4,5,10]

Table 7.10: Results of PLS-DA classification of FTIR spectra from grass pollen. 4 latent variables were used. The results are based on an independent test set of 600 spectra measured at different times.

affiliation / identified by PLS-DA as	*A. odoratum*	*B. inermis*	*H. bulbosum*	*L. perenne*	*P. alpina*
A. odoratum	72	46	55	34	17
B. inermis	5	44	46	24	1
H. bulbosum	2	8	3	7	3
L. perenne	3	11	13	21	1
P. alpina	37	11	5	34	97
Success rates	61 %	37 %	2 %	18 %	82 %
Overall success rate	40 %				

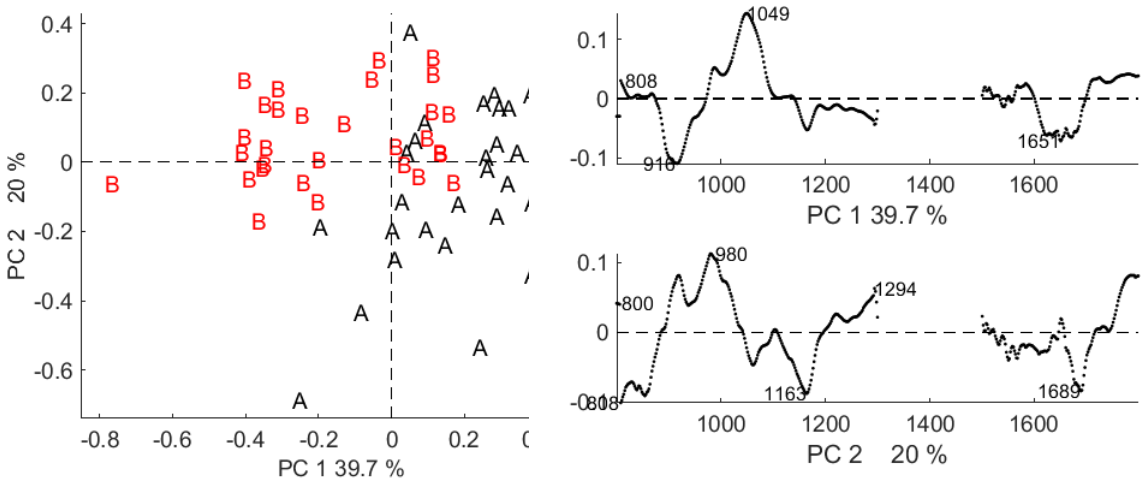

Figure 7.20: (Left) Scores plot and **(right)** loadings of the first and second PC for averaged pollen FTIR spectra measured at two different times A, black and B, red.

7.5.2 Classification in the presence of additional pollen species

To the 50 samples of the sample set Pollen Norway IIa, in total 21 samples from 4 different Poaceae pollen species (Pollen Norway IIb) were added tand FTIR and MALDI spectra were obtained. As discussed in the previous section of this chapter, some of the FTIR-spectra, particularly *Hordeum bulbosum* and *Bromus inermis* are not clustered with respect to the species by HCA, but rather cluster together according to their assignment to the tribe Triticodae (cf. Figure 7.14). By adding more data from different species, the variation regarding taxonomy is expected to change. It will be considered in more detail here.

In Table 7.11 the results of the PLS-DA model are shown conducting full-CV on 1423 pollen spectra from 71 individual plants (approx. 20 pollen grains for each plant). The pollen spectra from the embedded samples were pre-processed using the concatenated spectra from approach 2. The overall success rate is low with 47 %. Only spectra of *Poa alpina* show a high success rate of 89 %. For the two pollen samples *Hordeum vulgare* and *Piptochaetium avenaceum*, success under 5 % were determined.
Since the sample set shows a high variation in their taxonomical relationship, the misclassified spectra also have to be pointed out. For instance, 68 % of the spectra of *Anthoxanthum odoratum* are classified correctly. The majority of the misclassified spectra are identified as either *Lolium perenne* or *Poa alpina*. Similiar, *Lolium perenne* spectra are mostly misclassified as *Poa alpina* or *Anthoxanthum odoratum*. As discussed above the three pollen species are closely related to each other and are categorized into the same tribe Poaeae (Figure 7.19). It also should be pointed out, that a high percentage of spectra from other pollen species are misclassified as *Anthoxanthum odoratum*, *Lolium perenne*, and *Poa alpina*. This might be

caused by several unintended variances within the data set. First, the two different measurement times can affect the classification results greatly, as discussed above and second, the sample set of Pollen Norway IIb has a smaller amount of samples and therefore fewer spectra. The sample set Pollen Norway IIa consists of pollen from 10 individual plants, whereas Pollen Norway IIb comes from 5-6 individual plants.
A biased data set can cause a weighting of the variables towards the larger data sets. This problem is highlighted in literature, and studies suggest procedures to overcome this bias.[227,228] Another promising concept is a hierarchically ordered taxonomic classification by partial least squares (Hot-PLS), where the taxonomical information can be used to improve a model.[229] For comparison, PLS-DA was also executed using 70 mass spectra from the sample set Pollen Norway II (a+b).
As in the FTIR experiment, the mass spectra for Pollen Norway IIa were obtained on a different measurement day than Pollen Norway IIb. Also, for sample set Pollen Norway IIb, a different MALDI-TOF device was used. Changing measurement time and device leads to another source of variance, similar to the variance in the data set of the FTIR microspectra..
The amount of correctly classified and misclassified spectra, as well as the success rates, are presented in Table 7.12. The overall success rate is over 80 % for the classification of nine different pollen species from the different pollen tribes. Mass spectra of *Anthoxanthum odoratum*, *Bromus inermis*, *Hordeum bulbosum*, *Poa alpina*, and *Piptochaetium avenaceum* are 100 % correctly assigned to their respective pollen species. 2 spectra of *Lolium perenne* are misclassified as *Bromus inermis*. Interestingly, *Lolium perenne* is not as close as related to *Bromus inermis* than to e.g. *Poa alpina* and *Anthoxanthum odoratum*, this misclassification does not fit the taxonomical relation (Figure 7.19). The two pollen species *Hystrix patula* and *Piptatherum millaceum* have both no correctly classified spectra. Spectra from *Hystrix patula* are mostly misclassified as *Hordeum bulbosum*, which is closely related to *Hystrix patula* as both are from the same plant tribe Triticeae. Likewise, the mass spectra from *Piptatherum millaceum* are completely misclassified as *Piptochaetium avenaceum*. The reason for misclassification can be explained by the taxonomical relationship. Both, *Piptatherum millaceum*, and *Piptochaetium avenaceum*, are plants from the same plant tribe Stipeae (Figure 7.19).
Overall, the classification results in Table 7.12 confirm the high reproducibility of MALDI-TOF mass spectrometry. From 70 samples only 2 spectra were misclassified, due to unknown sources of variance , and other misclassification are possibly due to the specific taxonomical relationships of the respective species. Considering the different amounts of spectra for each pollen species of measurements on different days and different devices MALDI-TOF MS in combination with chemometric methods is a suitable tool for automatic pollen identification.

Table 7.11: Results of PLS-DA classification of 1423 FTIR spectra from nine grass pollen species. 6 latent variables were used. The results are based on full cross-validation.

affiliation / identified by PLS-DA as	*A. odoratum*	*B. inermis*	*H. bulbasum*	*H. vulgare*	*L. perenne*	*P. alpina*	*P. avenaceum*	*H. patula*	*P. milla*
A. odoratum	135	19	25	20	47	7	7	6	11
B. inermis	13	98	25	12	25	7	7	12	3
H. bulbasum	10	34	131	21	28	9	8	3	4
H. vulgare	0	0	0	4	1	0	1	5	2
L. perenne	20	12	13	4	42	0	1	1	1
P. alpina	21	30	15	2	43	177	24	32	28
P. avenaceum	0	0	0	0	0	0	0	1	1
H. patula	0	4	0	28	8	0	28	51	22
P. miliaceum	0	2	0	9	2	0	23	9	29
Suceess rates	68 %	49 %	63 %	4 %	21%	89 %	0 %	43 %	29 %
Overall success rate	47 %								

Table 7.12: Results of PLS-DA classification of 70 mass spectra from nine grass pollen species. 6 latent variables were used. The results are based on full cross-validation.

identified by PLS-DA as \ affiliation	*A. odoratum*	*B. inermis*	*H. bulbasum*	*H. vulgare*	*L. perenne*	*P. alpina*	*P. avenaceum*	*H. patula*	*P. miliaceum.*
A. odoratum	10	0	0	0	0	0	0	0	0
B. inermis	0	10	0	0	2	0	0	0	0
H. bulbasum	0	0	10	0	0	0	0	5	0
H. vulgare	0	0	0	5	0	0	0	0	0
L. perenne	0	0	0	0	8	0	0	1	0
P. alpina	0	0	0	0	0	9	0	0	0
P. avenaceum	0	0	0	0	0	0	5	0	5
H. patula	0	0	0	0	0	0	0	0	0
P. miliaceum	0	0	0	0	0	0	0	0	0
Success rates	100 %	100 %	100 %	100 %	80 %	100 %	100 %	0 %	0 %
Overall success rate	81 %								

In summary, to avoid the scattering artifacts in single pollen spectra, pollen were embedded in paraffin. The scattering is reduced in spectra from embedded pollen grains compared to spectra from un-embedded pollen grains and spectral variances within the same sample are reduced. Different pre-processing approaches were applied before data analysis to minimize the paraffin contribution. The spectra were classified using PLS-DA and other machine learning approaches, indicating that high success rates are dependent on the pollen species that need to be classified. The misclassified spectra are often classified as pollen spectra from grass species that are closely related to each other.
It was shown here that the FTIR measurements of single pollen grains suffer from relatively low reproducibility. Specifically, the results have shown that the classification model of FTIR spectra is only reliable when training and test sets are measured within a short time. Nevertheless, it i would be expected that classification approaches may lead to better performance, when the training set includes spectra of measurements from several different times and conditions. The addition of a new data set that was measured at a different time, to an existing data set did not lead to successful discrimination of an independent test set, here. On the contrary, MALDI TOF MS experiments on the extracts from many pollen grains showed higher reproducibility. This indicates. that measurements on single pollen grains (regarded as biological replicates) introduce an additional source of variation. Nevertheless, the utilization of single pollen grain FTIR spectra in multiblock analysis together with other spectroscopies, or the use of the microspectra together with bulk FTIR data[10] could be very promising.

8 Classification of MALDI MS images of different pollen species in mixtures

To utilize matrix-assisted laser desorption/ionization (MALDI) mass spectra for automated identification, both the experiment and the data analysis, need to be optimized. The chapter discusses the accessibility to detect different pollen species simultaneously in MALDI mass spectrometry images (MSI) of pollen mixtures by applying partial least square-discriminant analysis (PLS-DA), artificial neural networks (ANN), random forest (RF), and non-negative matrix factorization (NMF).
Before a MALDI MS experiment, the pollen grains are usually treated with formic acid and the molecules are extracted by the acid.[1,2,107] Extracts from different pollen species can overlap on the target, and therefore several peaks in the obtained MALDI mass spectra might be suppressed or even masked.[57,230] As will be discussed, these aspects also became evident in the data analysis.
All MALDI-TOF MS experiments, which involves the acquisition of the reference spectra and the MS images, were optimized and performed by Dr. Franziska Lauer.[57]

8.1 Classification of MALDI MS images of pollen in mixtures

A mixture consisting three pollen species of unknown composition is spread over a certain area on a carbon tape and measured using MALDI TOF MS with 100 μm distance between to sampling points. An image of such a mixture from the three species *Artemisia absinthium*, *Betula occidentalis*, and *Populus nigra* is analyzed using different classification approaches. Figure 8.1 shows the pollen, fixed on the carbon tape Figure 8.1 (left), as well as the polygonal scanning geometry of the measurement Figure 8.1(right).

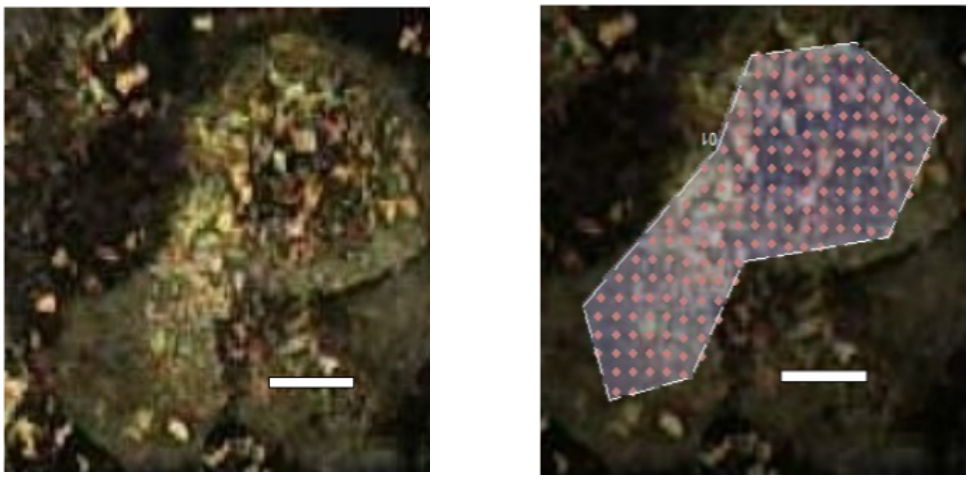

Figure 8.1: (Left) Bright-field images of the pollen on carbon tape and corresponding polygonal measurement geometry. **((Right)** Each red dot represents a represents a sampled spot, scale bar, 500 μm. Micrographs were taken by Dr. Fransiska Lauer.

The analysis of a mass spectral image addresses several challenges that can be compared to problems in real field studies:
(I) The amount and distribution for the pollen of each species are unknown. One of the main drawbacks in MALDI TOF MS is the acquisition of spectra from an acid extract, rather than from a pollen grain. Therefore, the results of classification cannot be validated for images. Furthermore, sizes and shapes of the pollen grains from the investigated samples can differ and therefore the composition of these extracts is unknown as well.
(II) Pollen is spread on carbon tape. Previous results have shown, that a species-discrimination of pollen is also possible using sticky carbon tape for fixation.[107] Nevertheless, the quality of the spectra obtained from pollen on carbon tape might be lower than in the case of the steal-less target used for the classifications of MALDI mass spectra in the previous chapters.
(III) MALDI TOF MS enables the investigation of a sample with a certain depth. Overlap and overlay of the pollen extracts can mask spectra from underlying analytes.
(IV) A suitable geometry has to be chosen to minimize the duration of the laser shots. In automatic MALDI experiments, it is challenging to avoid bad spectra, particularly because of the lower quality in artificial pollen mixtures, due to suppressing effects of the pollen.[57]
Several classifiers can be applied to characterize such an image data set. In the following, the classification of the MSI is executed using (i) HCA, (ii) PLS-DA, (iii) ANN, and (iv) RF. The aim is to compare and evaluate the results of the different approaches and to discuss to which extent they could be used for an automatic pollen identification.

A classification using HCA differs from the other methods, since it is not based on a training set, but rather relies on the comparison between the reference and clustered spectra.[55]
In Figure 8.2, A, the pre-processed and averaged MALDI-Spectra in the range *m/z* 5000- 9000 are presented. The three investigated pollen species *Artemisia absinthium* (32 reference spec-

tra), *Betula occidentalis* (30 reference spectra), and *Populus nigra* (32 reference spectra) show species-specific spectra, that enable a clear differentiation of the spectra by eye. Nevertheless, difficulties occur in pollen mixtures, due to the experimental conditions. The extracts from the different pollen species may overlap and as a result, all of the mixture spectra here contain the dominant peaks from *Populus nigra* at *m/z* 6350 (Figure 8.2, B). Due to the specific pattern of the mixture spectra, all spectra would have been assigned as *Populus nigra* by peak picking. Nevertheless, the averaged spectra of each cluster (Figure 8.2, B) show some features that can be compared to the spectra from the reference measurements of the three pollen species. *Populus nigra* spectra have broad peaks between *m/z* 7200 and 7700 with the maximum at *m/z* 7570. Besides, two neighboring peaks at 8336 and 8500 are specific for *Populus nigra* regarding this sample set and are helping to discriminate *Populus nigra* spectra from the other spectra. All mentioned peaks (*m/z* 6350, 7570, 8336, and 8500) are most dominant in the spectra of cluster 2 (red). Therefore, the HCA analysis is suitable to identify *Populus nigra* spectra in this pollen mixture. The HCA image is helping to understand, how the *Populus nigra* pollen grains are most likely distributed on the target. Cluster 1 and Cluster 3 are more difficult to assign to one specific pollen species. The averaged spectra of Cluster 1 and Cluster 3 only slightly different peaks in the *m/z* range of 5000-5500. Therefore, according to peak similarities, cluster 1 could be assigned to *Artemisia absinthium* but Cluster 3 hardly shows the same peaks as *Betula occidentalis*.
(Figure 8.2, C) shows the dendrogram of the clustered spectra from the imaging data set. A threshold of 3 clusters can be chosen to color the respective branches in the dendrogram. Most of the spectra from the image would be sorted into Cluster 2 (Figure 8.2, C, red), which could be assigned to *Populus nigra*. Cluster 1 has the lowest amount of spectra (Figure 8.2, C, black), which forms a superordinate cluster together with Cluster 3 (Figure 8.2, C, green), due to their similarities discussed above.
In Figure 8.2 D, the results of the HCA classification are presented. Each spot is assigned to one of the three clusters and marked in the corresponding color. The middle part of the images is clustered as Cluster 2 and therefore assigned to *Populus nigra* (Figure 8.2 D, red). Spectra from Cluster 1 (Figure 8.2 D, black) are more located to the right border of the image, whereas spectra from Cluster 3 are located to the left and right border of the image (Figure 8.2 D, green).

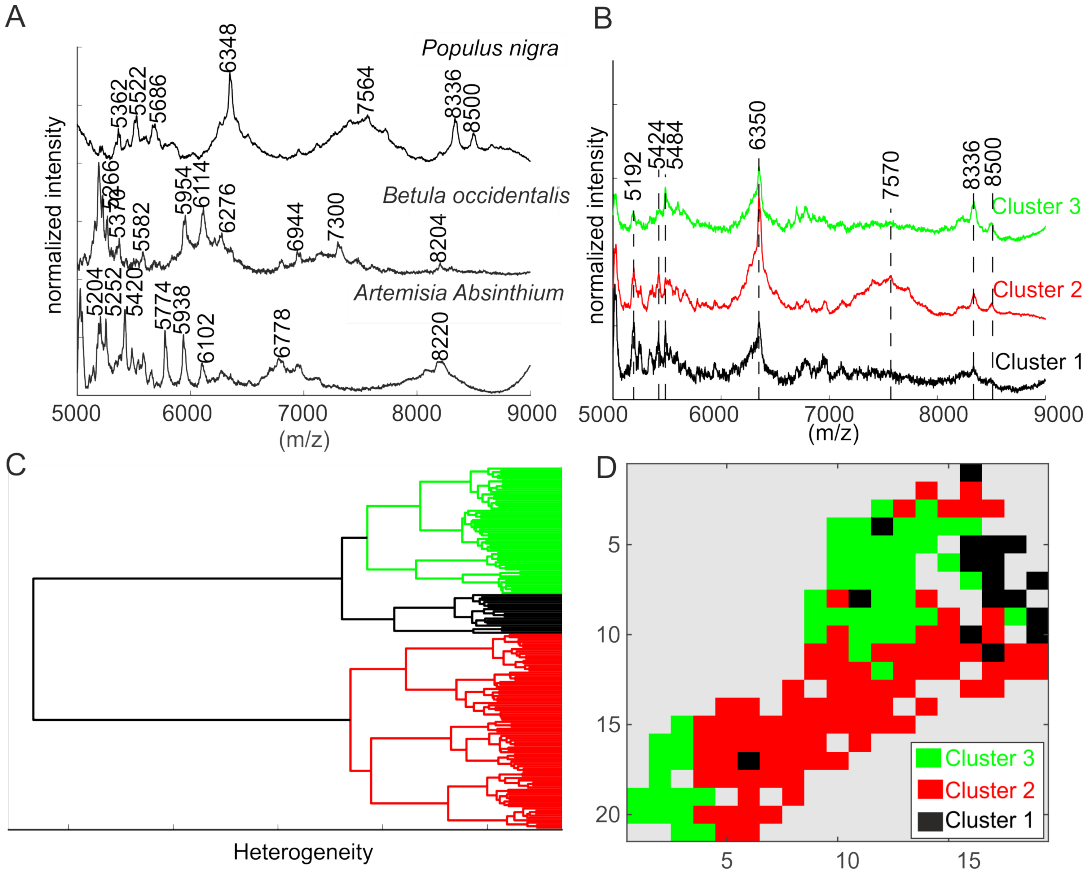

Figure 8.2: Identification of pollen species in a MS imaging data set using HCA.. **(A)** Averaged spectra of the three pollen species *Artemesia Absinthium*, *Betula occidentalis*, and *Populus nigra*. **(B)** Averaged spectra from each of the three main clusters obtained by HCA of the 154 spectra from the image. **(C)** Dendrogram of the 154 spectra of the image colored by clusters. **(D)** Clustering results of the image spectra. Cluster analysis was performed on 154 spectra using the spectral range *m/z* 5000- 9000 as input and *Ward's* algorithm for clustering.

HCA would be suitable for identifying pollen species with dominant peaks pattern in their spectra. Here, HCA is revealed to be more powerful than the analysis of individual peaks in the MALDI spectrum that was reported previously.,[56] where all image spectra would only be identified as *Populus nigra* as has also been reported recently by Lauer *et al.*[55] However, the interpretation of the HCA classification result requires information regarding the number of classes, that is, in this case of species to be discriminated. The clusters need to be evaluated by eye, which is not beneficial for the automatic identification of pollen. Also, the amount of clusters needs to be assessed as well. Figure 8.2 C indicates that the chosen number of clusters may not cover all variances of interest.

In contrast, unknown spectra can be identified using a chemometric model that is trained with spectra of known identity. The image of the unknown composition of the extract from the three different pollen species *Artemisia absinthium* (32 reference spectra), *Betula occidentalis* (30 reference spectra), and *Populus nigra* (32 reference spectra) is analyzed using a

model based on PLS-DA, ANN, and RF. All three methods are based on different theoretical backgrounds. PLS is based on the NIPALS (see Chapter 2) algorithm[61,67] and uses linear components to determine the classification, and ANN is trained with nonlinear components.[202] Furthermore, RF is a decision tree method, where random variables are chosen to build a tree.[76] The three approaches share the need to be calibrated on a training data set. The reference spectra displayed in Figure 8.2 A, were used for calibration.
Table 8.1 shows the validation results of the obtained classifiers of PLS-DA (Table 8.1, upper part) and ANN (Table 8.1, lower part). The validation of the PLS-DA model was executed using full cross-validation (full CV) and 3 latent variables. The model shows high success rates with an overall success rate of 98 %. 2 spectra of *Artemisia absinthium* were misclassified as *Betula occidentalis* (Table 8.1, upper part). In terms of classification, this is a reliable model for the prediction of the three species.

Table 8.1: Classification of the pollen species using PLS-DA and ANN. Classification results of PLS-DA (upper part) using full-CV and 3 latent variables and internal validation of the training set using ANN (lower part) using 2001 input units and 50 hidden neurons in the hidden layer.

affiliation / identified as	*Artemisia absinthium*	*Betula occidentalis*	*Populus nigra*
PLS-DA			
Artemisia absinthium	30 (93.75 %)	0 (0 %)	0 (0 %)
Betula occidentalis	2 (6.25 %)	30 (100 %)	0 (0 %)
Populus nigra	0 (0 %)	0 (0 %)	32 (100 %)
ANN			
Artemisia absinthium	32 (100 %)	0 (0 %)	0 (0 %)
Betula occidentalis	0 (0 %)	30 (100 %)	0 (0 %)
Populus nigra	0 (0 %)	0 (0 %)	32 (100 %)

For the evaluation of the ANN model, internal validation is applied so that the spectra used for the model are also tested on the calculated network. All spectra were classified as their affiliated species (Table 8.1, lower part), which indicates that the ANN model is suitable for the classification of samples of either one of the three pollen species.

Subsequently, the model is applied to the spectra of the MALDI MS image. Each spectrum is assigned to exactly one pollen species by the *winner-takes-all* approach.[202] The classification results can be visualized by presenting the MSI with the respective coloring for the species assignment (Figure 8.3). The classification results of the PLS-DA model is shown in Figure 8.3, left. Similar to the HCA image in Figure 8.2 D (red), the middle part of the image represents spectra that are different from the spectra of both borders. The spectra marked in blue are classified as *Populus nigra*, which is in good agreement with the outcome of the HCA (Fig-

ure 8.2). The spectra from the borders are classified as spectra from *Artemisia absinthium* (Figure 8.3, left, black), which shows similarities with the location and distributions of the spectra clustered in Cluster 1 and Cluster 3 in the case of the HCA image (Figure 8.2 D (green and black)). Also, in agreement with the HCA approach discussed above, almost no spectra from *Betula occidentalis* can be identified in the image.

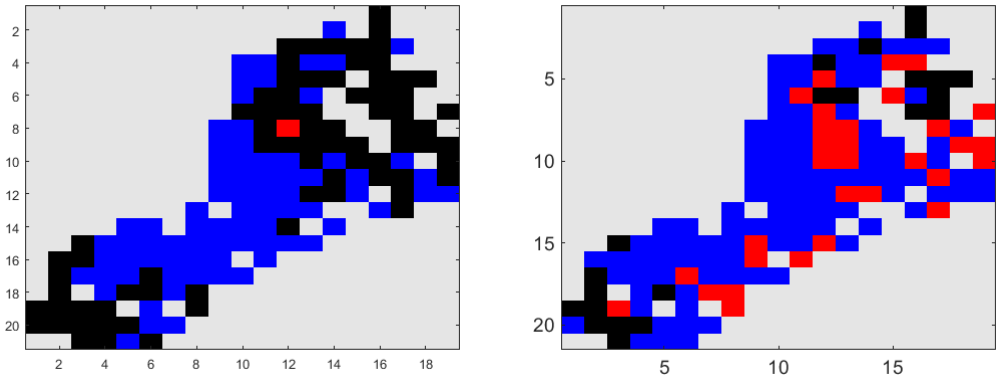

Figure 8.3: Representation of the classification results of the MSI of a pollen mixture containing the three pollen species *Artemisia absinthium*, black, *Betula occidentalis*, red and *Populus nigra*, blue using **(left)** PLS-DA, and **(right)** ANN.

The classification results of the ANN model are presented in (Figure 8.3, right). Most spectra are assigned to *Populus nigra* and the image shows similarities with the HCA (Figure 8.2 D) and PLS-DA images (Figure 8.3, left). In contrast to the PLS-DA image (Figure 8.3, left), more spectra distributed over the whole image, would be assigned to *Betula occidentalis*, instead of being classified as *Artemisia absinthium* (Figure 8.3, left, black). Most of these spectra (Figure 8.3, right, red) were clustered in Cluster 3 in the HCA, which indicates low reliability in the classification of the border spectra.

In addition to HCA, PLS-DA, and ANN, Random forest was applied to calculate a classifier and determine the spectra of the image from the pollen mixture. In Figure 8.4 (left), the decision tree for the training of the three different pollen species is drawn. On each node, the model adds branches of possible discrimination based on random variables until all spectra are branched into one definite species. In this example, the reference set consists of 94 spectra and 3 pollen species. The tree would first divide the set into two branches, one with *Artemisia absinthium* and *Betula occidentalis* and one with all three species. The latter would be split into *Populus nigra* and the two other species, that are separated in a third branch.
The corresponding classification image, based on *winner-takes-all* (WTA), is shown in Figure 8.4, right. Almost all spectra of the image are assigned to *Populus nigra* (Figure 8.4, right, blue). To some degree, this is in agreement with the classification results discussed above.

Nevertheless, only a few spectra at the right border are assigned to *Artemisia absinthium* (Figure 8.4, right, black) and only one spectrum was assigned to *Betula occidentalis* (Figure 8.4, right, red).

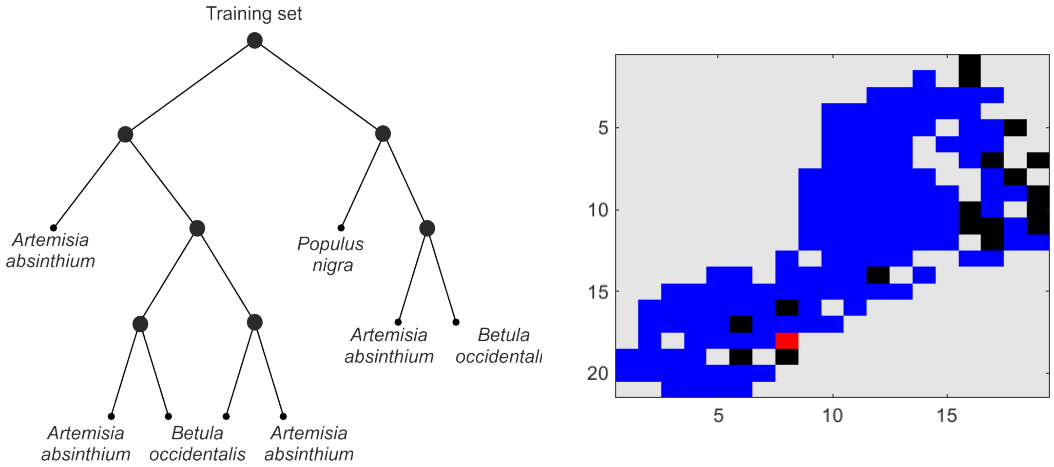

Figure 8.4: (Left) Random forest decision tree and **(right)** representation of the classification results of the MSI of a pollen mixture containing the three pollen species *Artemisia absinthium*, black, *Betula occidentalis*, red, and *Populus nigra*, blue. Random forest was trained using 500 trees.

Based on these assessments, the classification of unknown compositions of a mixture of pollen can be executed using different chemometric models.For the mixture applied in this example, all methods gave similar results and classified most of the spectra from the MSI as the one species *Populus nigra* with the most remarkable pattern in their reference spectra. In terms of automatic identification, supervised learning approaches are promising tools but are limited to the particular MALDI experiment, specifically with respect to number and combination of pollen species as well as the preparation.[57]

The total amount and real distributions of the pollen on the MALDI target remain to be detrmined. Furthermore, using a model based on the training with only a certain set of species in the mixture is not applicable in studies of real pollen samples. The usage of databases is essential in biotyping and well-designed for the classification of bacteria and fungi.[94,231–233] Here, a PLS-DA model based on a database of 2192 reference spectra comprising 16 different pollen species was trained to identify the three pollen species within the mixture.
In Table 8.2 the success rates for each species are summarized. The overall success rate using full-CV is 72 %, caused by some pollen species being correctly classified and others not. For *Artemisia absinthium*, the success rate is 0 % with misclassified spectra as *Betula tatewakiana* and *Pinus rigida*. In the case of *Betula occidentalis*, all spectra are also misclassified as *Betula tatewakiana* and *Pinus rigida*. For the third pollen species in the mixture, *Populus nigra*, a success rate of 59 % can be obtained with the misclassified spectra are assigned to *Betula tatewakiana* (misclassified spectra not shown).

Table 8.2: PLS-DA classification results of 2192 spectra from 16 different pollen species. 12 latent variables were used for calibration. The results are based on full-CV.

	Success rates [%]
Artemisia absinthium	0
Alnus cordata	78
Alnus rubra	0
Betula alleghaniensis	12
Betula ermanii	95
Betula occidentalis	0
Betula tatewakiana	88
Corylus avellana	77

	Success rates [%]
Corylus sieboldiana	100
Philadelphus californicus	72
Pinus mugo	0
Populus nigra	59
Philadelphus pubescens	100
Pinus rigida	100
Pinus sylvestris	0
Syringa reticulata	100

In the differentiation of the three samples in the MALDI mapping data sets, this PLS model indicates difficulties. The image based on the classification result obtained with the larger training set is presented in Figure 8.5 (left). Most of the spectra were identified as *Populus nigra*, which is in agreement with the classification approaches based on the reference spectra of three pollen species. (Figure 8.5 (left), blue pixels).
All green pixels correspond to pollen species that are not in the mixture and are thus misclassified spectra. The misclassified spectra are all identified as *Betula tatewakiana*. As discussed above the spectra of the three pollen species *Artemisia absinthium*, *Betula occidentalis*, and *Populus nigra*, are often misclassified as *Betula tatewakiana*.
The two pollen species *Artemisia absinthium* and *Betula occidentalis* cannot be identified in the mixture using the PLS-DA model with the database of 16 pollen species. This can be explained by either the poor predictability by the model with 0 % success rates each (Table 8.2 or, considering the HCA classification results above, by the masking of the dominant pattern of the *Populus nigra* spectra (cf. Figure 8.2, B), e.g., by suppression effects.[57,230]

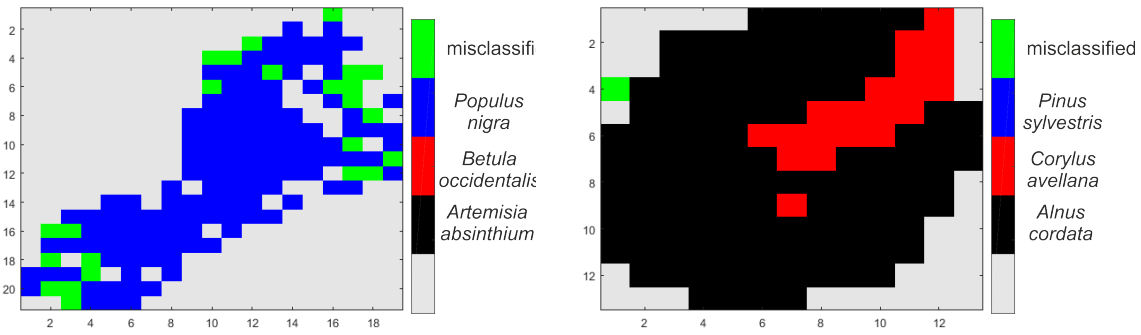

Figure 8.5: Classification results based on a PLS model of 16 different pollen species and 2192 spectra. **(Left)** Classification results of a mixture of *Artemisia absinthium*, black, *Betula occidentalis*, red and *Populus nigra*, blue. **(Right)** Classification results of a mixture of *Alnus cordata*, black, *Corylus avellana*, red and *Pinus sylvestris*, blue. Green pixels indicate misclassified spectra. Pixel size, 100 × 100 μm.

The model trained on the database with 16 pollen species can be applied to other compositions of different pollen species as well. Another image of the three different pollen species *Alnus cordata*, *Corylus avellana*, and *Pinus sylvestris* is investigated by PLS-DA (Figure 8.5, right). Using the model *Alnus cordata* and *Corylus avellana* can be discriminated with success rates of 78 % and 77 % in full-CV(Table 8.2). *Pinus sylvestris* cannot be described by the model (Table 8.1 and, in agreement with this, cannot be found in any imaging data set.
The classification results visualized in Figure 8.5 (right) state that most of the pixels are assigned to spectra from *Alnus cordata* (Figure 8.5, right, black). Moreover, red pixels are located in the middle part of the image that correspond to the spectra of *Corylus avellana* (Figure 8.5, right, red). Both pollen species are described by the model, so there is a high possibility, that the image of the pollen mixture of the pollen mixture indicates the presence of these two pollen species, which was indeed the case in this experiment.
When a random forest was trained based on the 16 pollen species and tested on the mixture of *Alnus cordata*, *Corylus avellana*, and *Pinus sylvestris* mixture with results similar to the PLS-DA analysis were obtained (data not shown). In good agreement with the results in Figure 8.5 (right), *Alnus cordata* and *Corylus avellana* are identified within the mixture, but no spectrum is assigned to *Pinus sylvestris*. The spatial distribution of the two identified species resembles that in Figure 8.5 (right), with slightly more spectra identified as *Corylus avellana* in the middle of the image.

The drawback in MALDI-TOF-MS is the extraction of molecules from many pollen grains prior to the measurements instead of probing single pollen grains. This leads to simultaneous probing of extract from the same and from different species. The loss of certain species-.specific patterns in such mixture spectra can be caused by peak suppression or by simple co-occurrence of different species-specific signals.[57] Using chemometric models based on pollen species in the mixture and based on a small database, dominant species with high specificity can be identified in mixtures, whereas those species whose reference spectra cannot be discriminated easily will not be identified within in the mixtures either. Therefore it is crucial to aim for a collection of very-high quality reference spectra, when using carbon tape as a substrate.[55,234,235]

8.2 Classification of pollen in mixtures using matrix factorization methods (NMF)

Due to the preparation step including extraction with formic acid, there is an overlap between different areas on the map, as mentioned before. Using the methods described above, the pollen can be identified, but their distribution on the target is questionable since all algo-

rithms follow the WTA principle. Mathematically, singular value decomposition is suitable for splitting up mapping data into their constituents. Since the factorization using PCA is weighted towards the variances, it is not well suited to find the right components, if there are many different pollen species and the variation is caused by different overlapping effects, e.g., suppression.[57] As it was shown previously,[71,236] and also by the analysis of the data in Chapter 7, NMF is better suited. It follows the idea of the Beer-Lambert Law, where the matrix can be factorized into components and relative contributions . The algorithm calculates iteratively the components, with the constraint of non-negativity.

No reference spectra are needed to factorize the image mapping data, but they are used to compare the calculated pure component with the averages of the reference spectra for the respective pollen species. Figure 8.6 shows the averaged spectra of the three three pollen species from the pollen mixture: *Alnus cordata*, *Corylus avellana*, and *Pinus sylvestris*.

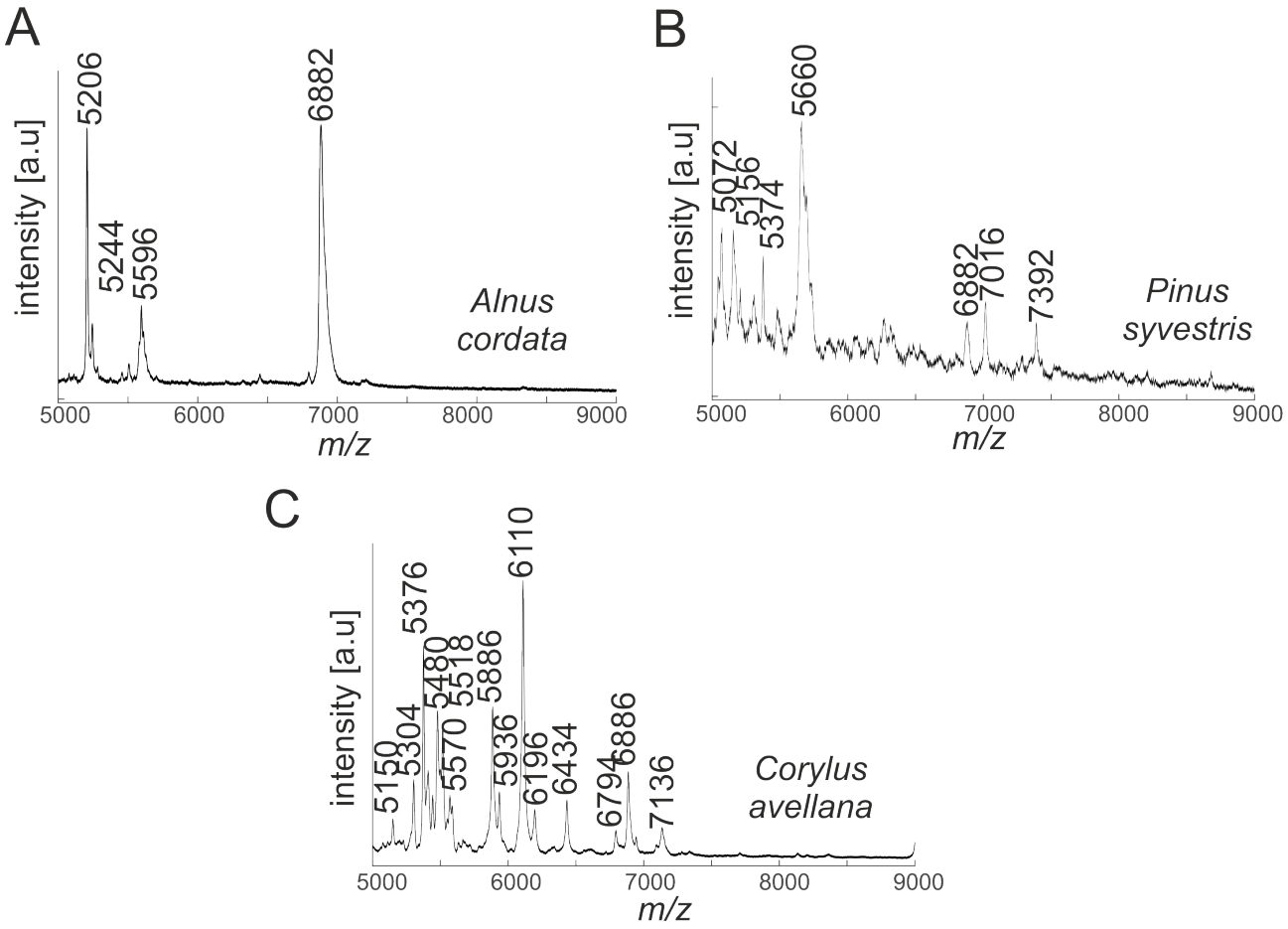

Figure 8.6: Pre-processed and averaged reference spectra of **(A)** *Alnus cordata*, **(B)** *Pinus sylvestris*, and **(C)** *Corylus avellana*.

NMF was conducted using the MSI of the mixture from the three pollen species. Figure 8.7 shows four pure components as well as the concentration profiles with the coordinates. In order to compensate for the low signal-to-noise ratio, four pure components were used rather than three, assuming that one pollen species could correspond to one component, and one

to the background.
Component 1 and its relative contribution to each pixel spectrum of the image are presented Figure 8.7 (A and B). The relative distribution of this specific component indicates that this pattern is found in the majority of the image spectra (Figure 8.7, A). The relative contribution of Component 1 is higher at the borders of the image and lower in the middle of the image. The component shows four dominant peaks at *m/z* 5206, 5244, 5596, and 6882 (Figure 8.7, B). The four peaks belong to the species-specific pattern of *Alnus cordata*. For comparison, the averaged spectrum of the reference spectra is shown as well in (Figure 8.7 (B), inset). Besides the peak position, also the ratios of the peaks of the average spectrum agree with those in Component 1, with *m/z* 5206 and 6882 have higher intensity compared to the peaks at *m/z* 5244 and 5596. It can be concluded that Component 1 represents the relative contribution of *Alnus cordata*.
In Component 2, two dominant peaks at *m/z* 5206 and 5596 are visible (Figure 8.7, D). This component shows a relative contribution at the upper border of the image(Figure 8.7, C). Still, Component 2 cannot be assigned to any of the three averaged spectra (compare with Figure 8.6). The two peaks at *m/z* 5206 and 5596 can be found in the spectra of *Alnus cordata* (Figure 8.7 (B), inset), but the dominant peak at *m/z* 6882 is missing in the pattern of Component 2. It is therefore very likely that Component 2 also describes and important part of the spectral pattern of *Alnus cordata* (Figure 8.6 (A)). The component could be incomplete because of suppression effects[57] in the mixture spectra or due to the factorization itself, which is not specifically surprising.
Component 3 shows high relative contributions in the middle of the image with the maximum in the middle and a lower relative contribution in the pixel spectra surrounding the maximum (Figure 8.7, E). Component 3 shows a relative specific peak pattern, Its peak pattern is similar to that in the average reference spectrum of *Corylus avellana* (compare Figure 8.7 (F) with Figure 8.6 (C)). Most of the peak positions and ratios of the peaks are similar in Component 3 and the average spectrum of *Corylus avellana*. On the contrary, the one peak at *m/z* 5206 in Component 3 cannot be found in the average spectrum of *Corylus avellana*, but in the average spectrum of *Alnus cordata* (Figure 8.7, B). Nevertheless, due to the high similarity of Component 3 with the average spectrum of *Corylus avellana*, it can be concluded, that Component 3 is a good indicator for the distribution of *Corylus avellana* pollen on the substrate.
Component 4 shows low relative contributions within the imaging data (Figure 8.7, G). It should be pointed out, that the color scale has its maximum at 0.6 of relative contribution, for easy visualization. Component 4 (Figure 8.7, G) shows peaks, that can be assigned to *Alnus cordata* (peaks at *m/z* 5206 and 5244 (Figure 8.7, B)) as well as *Corylus avellana* (peaks at *m/z* 5886, 5936, 6110 and 6882(Figure 8.7, F). Like Component 2, Component 4 cannot be assigned to one particular reference spectrum of any of the individual pollen species.
The identification and distribution of two of the three pollen species here can be revealed using NMF as a factorization. Overall, the results of the factorization are similar to these of

the PLS-DA, confirming that the middle part can be detected as *Corylus avellana* and the surroundings as *Alnus cordata* (Figure 8.5). Compared to the approach discussed in section 8.1, the misclassification of spectra from spatially overlapping extracts, e.g, as a consequence of suppression, is diminished.

As summary of Chapter 8, the utilization of MALDI mass spectrometry imaging data for pollen identification was presented. Two different images of three different pollen species were analyzed using several chemometric models to (i) identify the three species in the mixture and to (ii) describe their distribution on the target. The identification of the pollen species in mixtures is only valid if the respective pollen species is well described by the model that is applied. Both imaging data sets contain some pollen species in their mixtures that that fail to be detected using the applied chemometric models. The performance of each model differs only slightly from each other, suggesting that each of the models could be used to predict the distribution of the species on the target.
Using HCA, the image spectra are clustered and can be compared to the averaged reference spectra. Due to the occurrence of very dominant peaks, background issues caused by specific sampling on carbon tape,[55] suppression artifacts, and other reasons, classification by HCA can lead to misinterpretation.
Models based on three different chemometric approaches, namely PLS-DA, ANN, and RF were applied to the image data to identify the spectra. All models gave similar results, confirming that one or two different pollen species can be identified in the mixture. Specifically, a data base comprising spectra from 16 different pollen species instead of only three improves identification accuracy, as was observed for the PLS-DA analysis.
The challenge of overlapping extracts on the target can be addressed using factorization of the spectra by NMF.

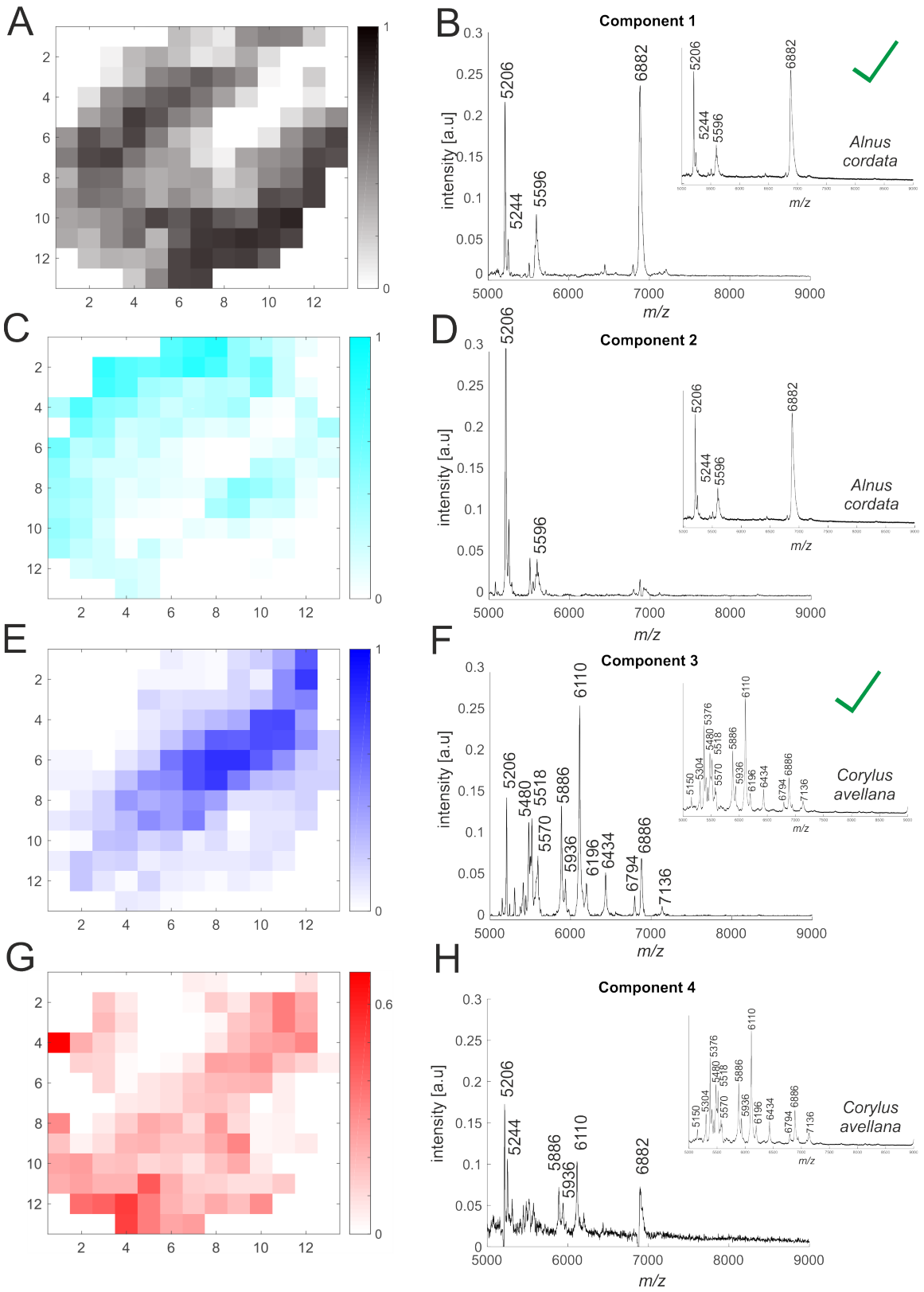

Figure 8.7: Relative contributions (**A, C, E, G**) and components (**B, D, F, H**) obtained by NMF of the MSI containing the three pollen species *Alnus cordata*, *Corylus avellana* and *Pinus sylvestris*. The factorization was conducted using 4 components. Pixel size, 100 × 100 μm.

9 Analysis of imaging data from plant tissue sections

Parts of the results presented in this chapter are published in: Multivariate Raman mapping for phenotypic characterization in plant tissue sections. Liedtke I*, Diehn S*, Heiner Z*, Seifert S, Obenaus S, Büttner C, Janina Kneipp. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*. 2021;251:119418.; doi: 10.1016/j.saa.2020.119418.
*Authors contributed equally to the work.

Raman microspectroscopy is a suitable tool for the investigation of complex biological tissues, as plant tissues and particularly Raman imaging is commonly applied to microscopic-sized samples.[6,7,152] A data cube with two spatial dimensions and the Raman-Shifts as a third dimension can be obtained. Therefore, each xy position consists of a full spectrum, which can generate data sets that have a high variance.
The Raman image of the tissue is usually visualized by a chemical image where each spectrum is represented by one key parameter, which reduces the data cube to a 2D map. The key parameter can be, e.g., the intensity of a certain position the area between a specific band and the baseline, or an intensity ratio between the key numbers of different bands.[7,33] As a result, the chemical image shows the distribution of such key parameters across all xy-positions and the varying chemical composition of the tissue can be detected and characterized.
In many research fields, such as botany or crop science, it is necessary to discuss the changes of the tissue composition due to some chemical treatment or environmental conditions. In these cases, small differences need to be described, and the investigation of only one chemical image would not be sufficient, but requires the comparison of data from many maps.
Due to the high heterogeneity within one Raman image, meaningful selection of spectra as well as suitable pre-processing steps are needed. In addition, the visualization of interpretable results from the Raman of classification experiments in the Raman images becomes crucial. In this chapter, the treatment of Raman imaging data is discussed. The variances within one data cube need to be reduced with respect to different biological questions. This is achieved using two different selection methods, univariate (based on, e.g., different intensities) and multivariate (hierarchical cluster analysis (HCA)). Specifically, multivariate analysis of the extracted spectra and the visualization of its results will be presented. Here, the selection and

analysis are executed using data sets from different projects:
I) The assessment of variation within in Raman mapping data of cross sections from *Cucumis sativus* Sonja in a sample set comprising different plant organs and two specific growth conditions.
II) The exploratory data analysis of one Raman map from a root tissue from *Sorghum bicolor* to identify and cluster map regions of a specific biochemical composition for selective data extraction.
III) The combination of spectroscopic data of cell wall compartments and additional plant information for the comprehensive characterization of Sorghum plants of different phenotype with respect to silica accumulation (*SbLsi*-mutant and wild-type). Data in this section were measured by Dr. Zsuzsanna Heiner, Ingrid Liedtke, Prof. Dr. Rivka Elbaum, and Nerya Zexer as part of collaboration projects*.

9.1 Selection of relevant spectra based on spatial information

During the Raman mapping measurements of thin cross sections from plant tissues, spectra are also obtained from areas with non - relevant information for the specific classification problem, e.g. the lumen regions or the empty regions of the CaF_2-slide. These sources of variance that add no further information, but cause problems in the investigation of the influence of small effects, and need to be discarded before further analyses. Moreover, the size of the data can be reduced drastically by the selection of relevant spectra in advance.

9.1.1 Univariate selection of cell wall spectra

The univariate approach to select spectra is useful for a fast, automatic reducing of high variances in the data cube caused by differences in the tissue's substructure. The method can be executed for a rather rough selection of spectra. The algorithm is based on the same principle as the chemical imaging. First, a key parameter such as the intensity or the integral of a specific band is calculated for each spectrum of the Raman map. As a result, one value for each xy-position is obtained. All values are min/max normalized so that each xy-position consists of a value between 0 and 1. A normalization of the key parameter is crucial to set a threshold between 0 and 1 that is valid for all maps in an automatic procedure.
The algorithm allows the selection of each Raman maps with adapting the threshold or all maps at once with the same threshold, depending on the quality of the data. In some cases,

* In order to remain consistent with potential other reports of the data (manuscript in preparation with several co-auhthors) and the anticipated PhD thesis of I. Liedtke (I. Liedtke, personal communication), all Raman Stokes shifts are assigned negative values in the plots of the spectral data.

the same threshold is not suitable for all maps. Particularly further consideration of the threshold is needed for irregularities like remaining spikes in the data, that can affect the key parameter.

In the following example, the selection is optimized for the discrimination of different plant organs from *Cucumis sativus* Sonja based on their cell wall spectra. To select the right key parameter, an exploration of the spectra is necessary. Figure 9.1 shows three representative single cell wall spectra from the three different plant organs stem, root, and leaf. Bands are obtained at 1098, 1141, 1378, 1336, 1598, and 1660 cm^{-1}. These bands can be assigned to lignin (1278, 1598, and 1660 cm^{-1}) and cellulose (1098, 1141, and 1378 cm^{-1}).[237] Lignin and cellulose are the main constituents in plant cell walls and would lead to an indication if a spectrum is assigned to the cell wall or to the lumen of the cells.

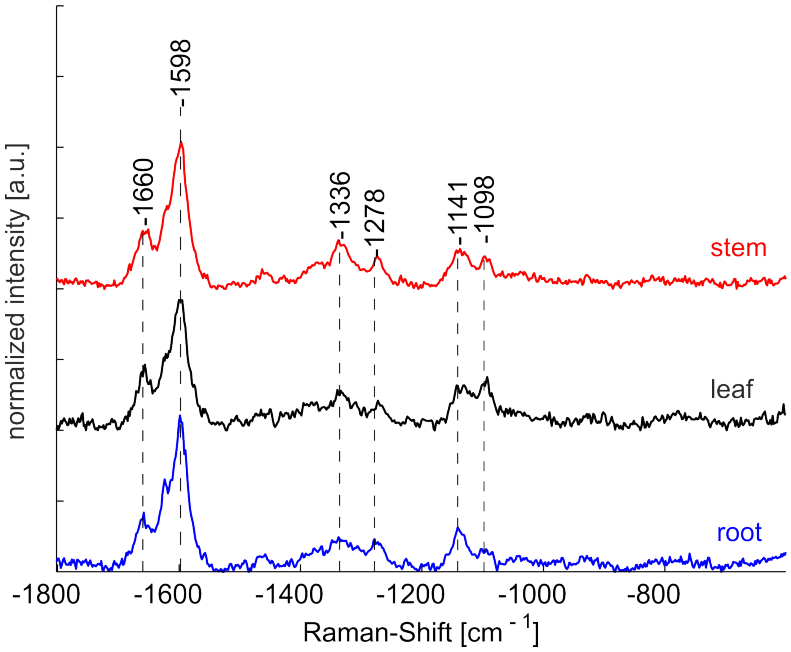

Figure 9.1: Representative single spectra of cross sections for three different plant organs stem (red), leaf (black), and root (blue) from *Cucumis sativus* Sonja (excitation wavelength, 532 nm, excitation intensity, $1.7 \cdot 10^6 W cm^{-2}$), accumulation time, 1 s. Spectra are pre-processed using interpolation, asymmetric least square (AsLS) baseline correction, and vector normalization as discussed in Chapter 3.

Figure 9.2 shows two chemical images (Figure 9.2 (A and C)) based on the integral of the spectral range between 1550 and 1700 cm^{-1} and the baseline. The differences in the chemical composition of these two maps (and other maps) need to be evaluated. Based on the distribution of the bands between 1550 and 1700 cm^{-1} that can be assigned to lignin,[50] both chemical images (Figure 9.2 (A and C)) show a high contrast between certain structures of the tissue. Spectra assigned to a key parameter of 0.3 and higher (Figure 9.2 (A and C), cyan to red) correspond to the plant cell wall. All spectra with a key parameter below 0.3 (Figure 9.2 (A and C (blue))) are obtained from the lumen and do not have relevant information regarding

the classification here and can be discarded.
In comparison, the chemical images in Figure 9.2 (A and C) indicate differences in their contrast of the cell wall regarding the lumen. Figure 9.2, C shows a more defined separation between lumen and cell wall spectra whereas in Figure 9.2, A the differentiation of lumen and cell wall spectra is less defined. In such a case the threshold for the selection of the cell wall spectra needs to be adapted to both Raman maps separately.
Figure 9.2 B and D) present the corresponding results of the selection of the cell wall spectra. A threshold of 0.35 (Figure 9.2, B) and 0.25 (Figure 9.2, D) are used, respectively. The selected spectra (Figure 9.2 (B and D), black) are those cell wall spectra from the chemical images with higher value for the selected key parameter (Figure 9.2, A and C (cyan to red)).

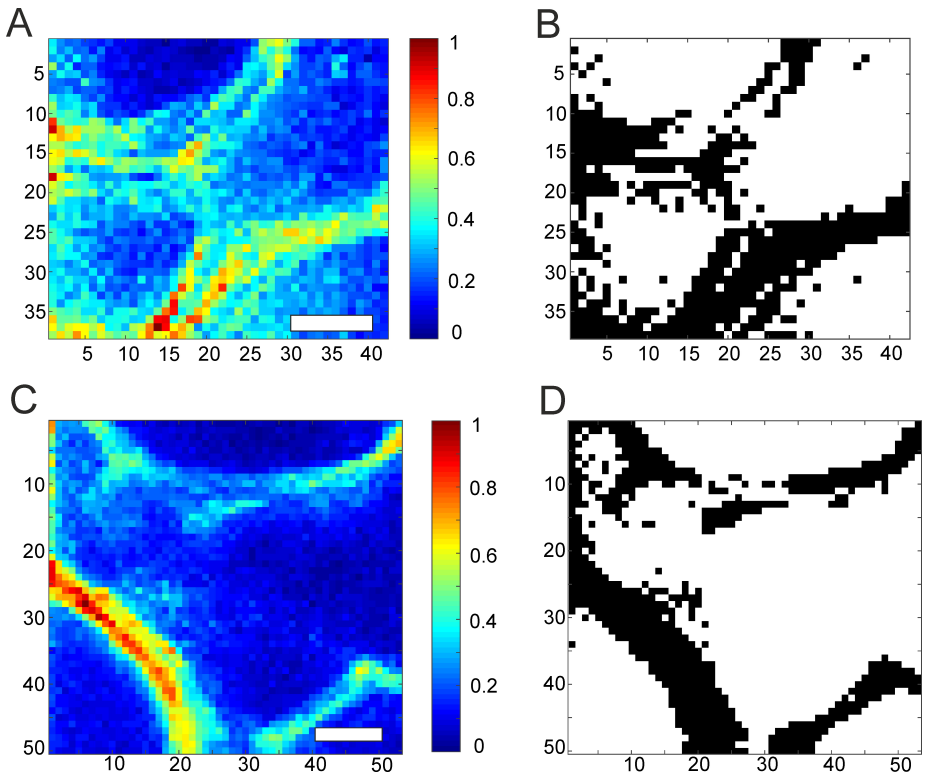

Figure 9.2: (**A** and **C**) Chemical images of two leaf cross sections from *Cucumis sativus* Sonja based on the integral of the spectral range between 1550 and 1700 cm^{-1} and the baseline of the baseline-corrected spectra and (**B** and **D**) the corresponding selected spectra (black) using a threshold of B, 0.35 and D, 0.25. Scale bar, 10 μm.

Spectra from the lumen regions could also be found in the selected spectra as well as cell wall spectra could be discarded. The selected spectra do not necessarily represent a meaningful data set, as and example, here, also spectra from the lumen of the cells were selected. This selection algorithm is a rough reduction of the variances in the data set and the data size, and is useful, when a complete and accurate separation of different substructures is not required.

For optimization of the data analysis, the selection/extraction algorithm may need to be more precise. In order to exclude/include only spectra with bands that indicate, e.g., a high content of cellulose from the data set, the algorithm can be extended. Figure 9.3 shows the chemical images of the two ranges between 1070 and 1108 cm^{-1} (Figure 9.3, A, cf. Figure 9.1 bands at 1098 and 1141 cm^{-1}), and 1313 and 1358 cm^{-1} (Figure 9.3, C, cf. Figure 9.1 band at 1336 cm^{-1}). These two bands can be associated with cellulose and also with other bands in the same regions such as lignin.[237]

A selection based on the integral of one of the two bands may include non-relevant spectra since the contrast between lumen and cell wall spectra is less enhanced for smaller bands. Figure 9.3, A and B represent the chemical image and corresponding selection based on the most prominent cellulose band between 1070 and 1108 cm^{-1}. The contrast in the chemical image (Figure 9.3, A) is high enough to discriminate between spectra, that can be assigned to cell wall spectra (Figure 9.3, B, black) and lumen spectra.

A selection of spectra based on a less dominant cellulose band 1313 and 1358 cm^{-1} (Figure 9.3, C) is less precise and the amount of lumen spectra within the selected data set would be higher (Figure 9.3, D, black), which leads to more non-relevant variances in the data set.

A more precise selection would be to increase the probability of the signal of the identification of a specific signal for mapping by using the product of two bands.[7] The product of the integrals of the two bands between 1070 and 1108 cm^{-1} and between 1313 and 1358 cm^{-1}. In the chemical image in Figure 9.3, E fewer spectra with a higher value for the key parameter (in this case the product of the two intensities) are present and therefore a selection of spectra with higher indication of cellulose takes place. The amount of selected spectra is smaller so that the upcoming data analysis can be focused on a smaller data set and unimportant heterogeneity is excluded.

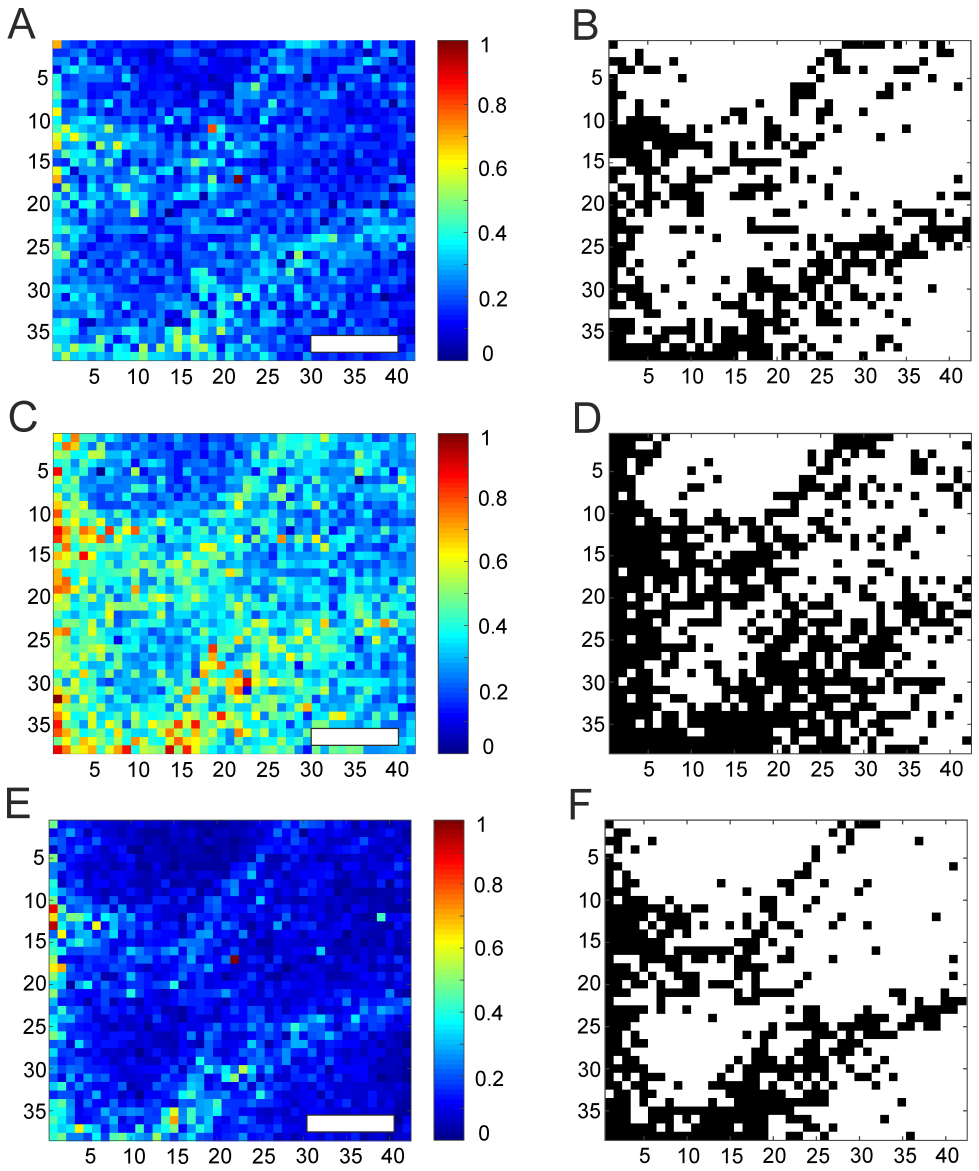

Figure 9.3: (**A, C**, and **E**) Chemical images of the same leaf cross section from *Cucumis sativus Sonja* as discussed in Figure 9.2 based on the integral of the bands between 1070 and 1108 cm^{-1} (**A**), the integral of the bands between 1313 and 1358 cm^{-1} (**C**), and the product between the two bands (**E**). (**B, D**, and **F**) the corresponding selected spectra (black) using a key parameter of **B**, 0.25, **D**, 0.35 and **F**, 0.15. Scale bar, 10 μm.

9.1.2 Multivariate selection of Raman spectra using HCA

An univariate selection based on a threshold of some key parameter of relevant spectra is a suitable easy tool for handling large data sets from Raman maps. Besides the reduction of the data size, a selection of a specific group of spectra is useful to analyze the differences between several maps, due to less non-relevant variances. For more accurate classification outcomes, a multivariate selection based on the HCA images of the cross sections can be more useful. Multivariate imaging methods as clustering or based on principal component analysis

(PCA) are commonly used for the analysis of Raman maps.[7,36,47,49,81,151,238] Regarding the clustering methods, the spectra were grouped and colored based on their spectral properties. In Figure 9.4 the clustering results of one *Sorghum bicolor* (wild-type) cross section are presented as another example of a Raman mapping data set. The HCA was executed using euclidean distances and *Ward's* algorithm.[58] The obtained dendrogram was divided into the three biggest clusters. Figure 9.4 (left), shows the classification results of the Raman map with the three colors representing the three different clusters. The positions of the clustered spectra of the three clusters can be marked using different colors in the HCA image. Figure 9.4 (left) suggests that spectra from the same clusters can be assigned to the same ultra-structure in the plant tissue. Spectra, that are in Cluster 1 (Figure 9.4, grey) represent the lumen region of the cross section, whereas Cluster 2 (Figure 9.4, black) and Cluster 3 (Figure 9.4, red) can be described as cell wall spectra. The amount of clusters can be chosen arbitrarily, yet *ad hoc* knowledge can be taken into consideration.

A plant cell wall consists of layers with a different composition of biopolymers, most abundant lignin and cellulose.[?] Taken into account that the step size of the Raman experiment here is 1 μm, it can be assumed that not all layers can be resolved. Nevertheless, the Clusters 3 (Figure 9.4, red) can be tentatively assigned to mainly the middle lamella, and Cluster 2 (Figure 9.4, black) is representing the layers in the border region of the cell wall.

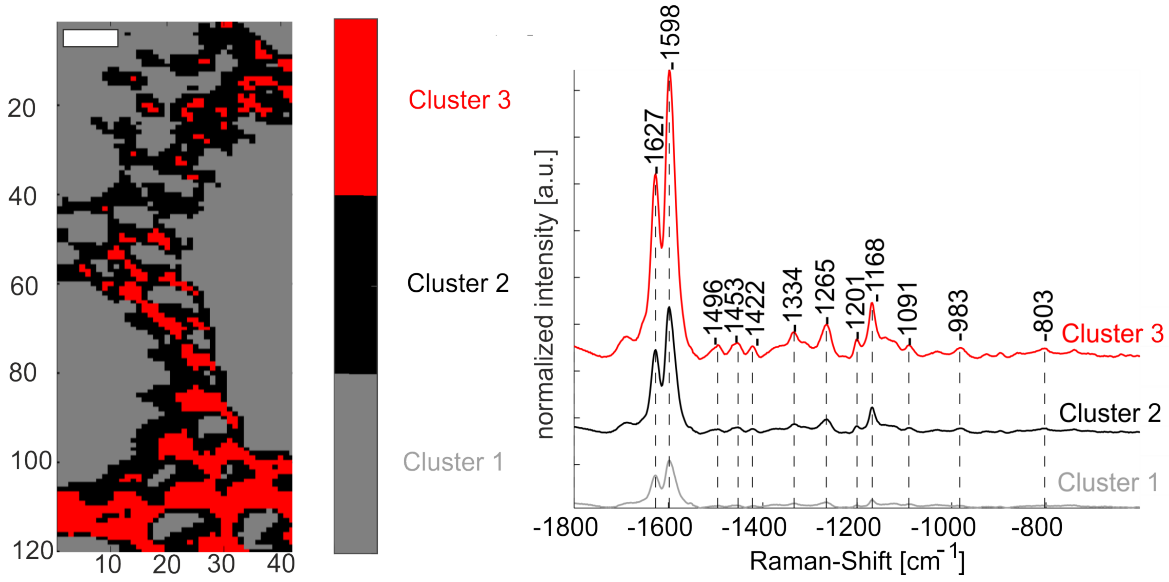

Figure 9.4: (Left) Classification results based on HCA and **right** averaged spectra for each cluster for one leaf cross section of *Sorghum bicolor* (wild-type) spectra were interpolated in the range of 600 - 1800 cm^{-1} proposed by Eilers[177] The spectra are stacked for clarity. HCA were executed using the pre-proccessed spectra and the Euclidean distances. In order to form the clusters, *Ward's* algorithm was applied. Scale bar, 10 μm

Figure 9.4, B shows the corresponding averages of the interpolated and baseline corrected spectra for each cluster. All spectra show bands that are specific for plant cell walls. The aver-

aged spectra for all three clusters show a high intensity of the bands at 1598 and 1627 cm^{-1} which can be assigned to the aryl ring stretching and the conjugated C=C stretching of lignin respectively.[239] The intensity is increasing from Cluster 1 to Cluster 3 (Figure 9.4, B). In a following analysis like PCA, the spectral differences of the three clusters would interfere as an additional source of variance within the data set. The clusters can be analyzed separately. The selection or omitting of all spectra from clusters can reduce the overall variances with respect to the biological question and simplify the analysis and interpretation of the results. Depending on the purpose of the data analysis, the selection of a specific structure of the tissue is required. Figure 9.5 gives an advanced example, where a selection of spectra from certain positions is needed. The data set obtained from a Sorghum root section was measured as part of a project by co-operation partners Prof. Dr. Rivka Elbaum and Nerya Zexer. The aim of the project is the characterization of silica deposition in *Sorghum bicolor* roots. The plants were cultivated in absence of silicic acid in order to assess the spots where silicification should occur.[40]

Figure 9.5, A shows the bright-field images of the endodermis. The spots of the Si-deposition are indiscernible in a common light microscope. A chemical image based on the integrals of the most prominent bands between 1550 - 1700 cm^{-1} (Figure 9.5, B) reveals two parallel rows of spherical areas with higher intensity (Figure 9.5, B, yellow and above).

In order to study the chemical composition of these spots, a refined selection is implemented. Using the HCA approach discussed above and presented in Figure 9.4 with five clusters the classification results presented in Figure 9.5, C are obtained. The spectra of the Si- deposition spots can be discriminated from most of the remaining spectra of the tissue and form two out of five clusters (Figure 9.5, C, red and green). Nevertheless, a large amount of spectra would also be included in these clusters, which can be seen in Figure 9.5, C on the left border of the map. Therefore, it can be concluded, that the number of clusters is not efficient for separating spectra from the Si-deposition spots and the remaining data set. One solution could be to increase the amount of clusters in order to have a more precise selection of clusters, yet the optimal number of clusters is unknown and can just be obtained by comparing the results of imaging with different numbers of clusters, i.e, 4 clusters, 5 clusters, 6 clusters, etc.

A more suitable approach is the extraction of those clusters that include spectra from the regions of interest in a first step and a subsequent calculation of a second dendrogram with this reduced data set.

In the example here (Figure 9.5, C), the spectra assigned to the Clusters 1, 4 and 5 (black, blue and cyan), were omitted from the data set. The spectra of the remaining Clusters 2 and 3 (Figure 9.5, C, red and green) are reassembled and afterwards separated again into five clusters using HCA (Figure 9.5, D). While the clustering of the spectra in Figure 9.5, C was based on interpolated and baseline corrected data, the spectra, used for the HCA image in Figure 9.5, D were vector normalized in addition to avoid variance sources e.g. based on different tissue thickness.

Cluster 3 and Cluster 5 in Figure 9.5, D (green and cyan) would represent the Si-deposition spots. It should be mentioned that pre-processing of the spectra in the different steps of this procedure can affect the outcome of this extraction. However, here, it is a very good approximation of the tissue region of interest.
Using a second cluster analysis, the selected spectra of this Raman map can be analyzed together with selected spectra from several other maps. In addition, the variances in the chemical composition of different structures within the same map can be compared. The selected and assigned spectra from one map can be renamed according to their clusters and analyzed. The naming of certain structures in Raman maps would be beneficial in order to interpret a following data analysis such as PCA and PCA imaging.

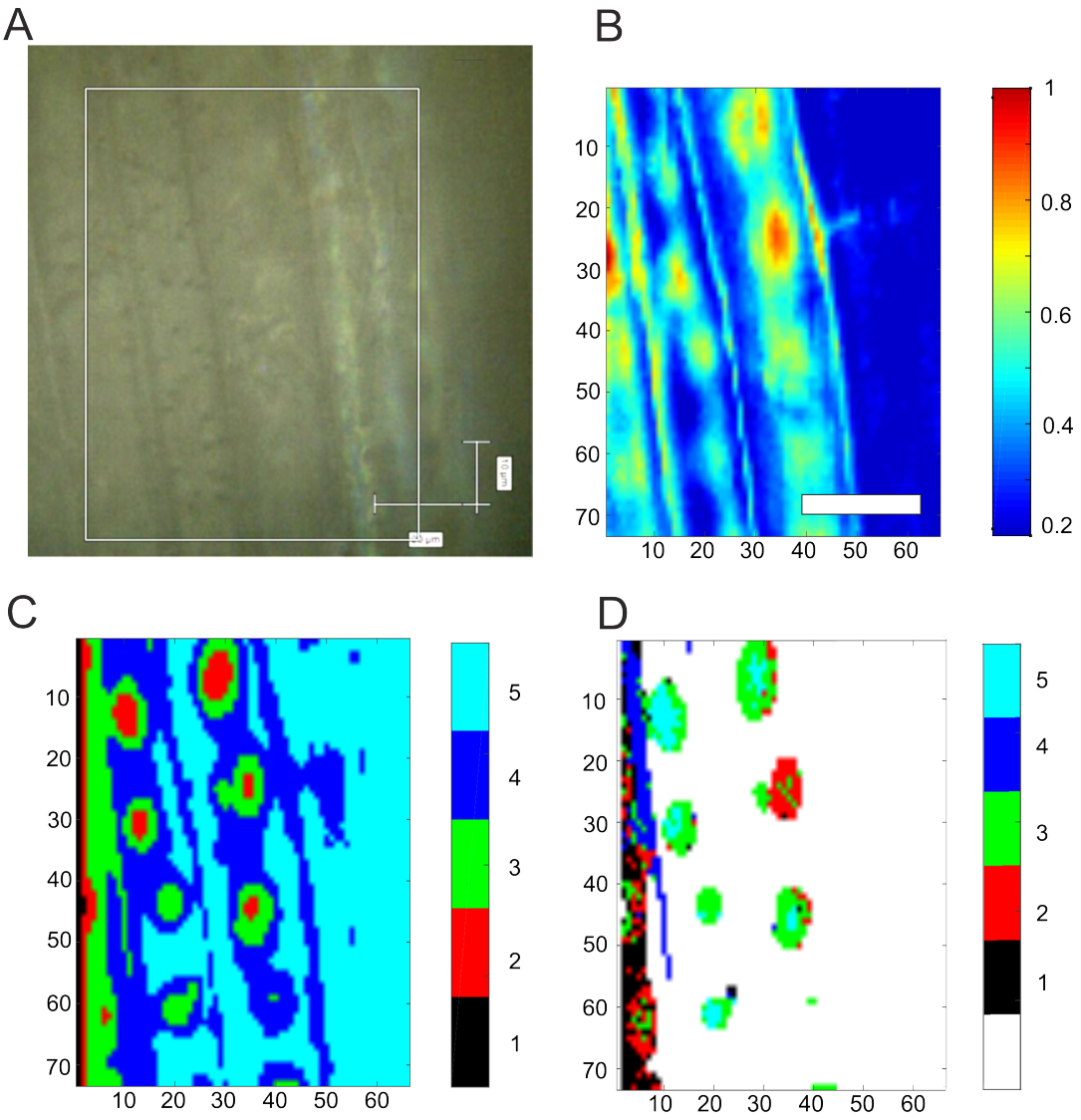

Figure 9.5: HCA clustering of a Raman image of root tissue and multivariate selection of spectra. (**A**) Bright-field image of root tissue. (**B**) Corresponding chemical image based on the integral of the spectral region between 1550 - 1700 cm^{-1} and the baseline. (**C** and **D**) HCA classification results of all 4818 spectra from the Raman map (**C**) and the clustering results of the HCA using spectra from the Clusters 1, 2 and 3 from C. (**D**) of the root section of *Sorghum bicolor* (wild-type) Spectra were interpolated in the range of 400 - 1800 cm^{-1} and baseline corrected. HCA were executed using the pre-proccessed spectra and the euclidean distances. In order to form the clusters, *Ward's* algorithm was applied. Scale bar, 20 μm.

9.2 Multivariate analysis of Raman mapping data

Since a large amount of spectra can be obtained from Raman imaging experiments, the interpretation of a following data analysis is challenging. Whereas the previous section deals with the reduction of variances, due to an adequate selection of spectra, in the following section the limitation of the data analysis and possible solutions will be discussed. Therefore

some examples of analyses of large data sets from various projects in cooperation with Ingrid Liedtke, Dr. Zsuzsanna Heiner, Prof. Dr. Rivka Elbaum, and Nerya Zexer will be presented.

9.2.1 2D histograms of score values from extracted Raman imaging spectra

As a first example, the Raman map of a root tissue discussed above, will be analyzed using PCA. Spectra assigned to the Si deposition spot (Cluster 3 and 5 in Figure 9.5, D) can be compared to spectra from the other regions of the root tissue, e.g., Cluster 4 from Figure 9.5 (C).
The scores plot in Figure 9.6 (A) shows overlapping symbols, due to the high amount of score values (371 spectra of the Si-deposition spots and 1541 spectra of Cluster 4). Several approaches to gain visibility in scores plots of big data are discussed in literature. Most prominent approaches are e.g. the clustering of score values,[240] the smoothing of the scatter plot,[241] and the visualization using binned scatter plots.[242] The latter approach is also adapted for the presentation of the scores plots here.
Plotting the score values separately for each group of score values is an appropriate solution for the overlapping issue. It leads to a better overview of the distribution of the score values on the scores plot. In addition, the 2D histogram can be generated from the score values. Besides the distribution of the score values, the maxima of these distributions can be localized.
A 2D histogram of the score values can be obtained by dividing the score plot into a certain amount of bins and counting the amount of score values of each bin. This can be executed by eg. the in-built Matlab-fuction *linespace* followed by *interp1*.
Figure 9.6 (B) shows the 2D histogram of all score values from the score plot presented in Figure 9.6, A. The histograms were obtained by a modification of the Matlab-exchange function *ndhist*, where the amount of bins are optimized automatically using Scott's normal reference rule.[243] The distribution is visualized using a customized color code, where a very low amount of score values/bin or no score value/bin is white, blue is presenting low amount of score values/bin and red the highest amount of score values/bin. The distribution of all score values shows one maximum.
Figure 9.6 (C and D) show the distributions for both groups separately. Both distributions have their maximum close to the origin of the scores plot, which indicates that the spectra of both groups, the spectra from the Si deposition spots (si, red) and other areas from the same tissue (co, black) have similar structure. It should be pointed out that most of the score values of spectra from the Si deposition spots have mostly positive values for PC 1 and PC 2 (Figure 9.6 (C)), whereas the score values from the other spectra are distributed over the scores plot of PC 1 and PC 2 (Figure 9.6 (D)). A following classification would lead to good results.

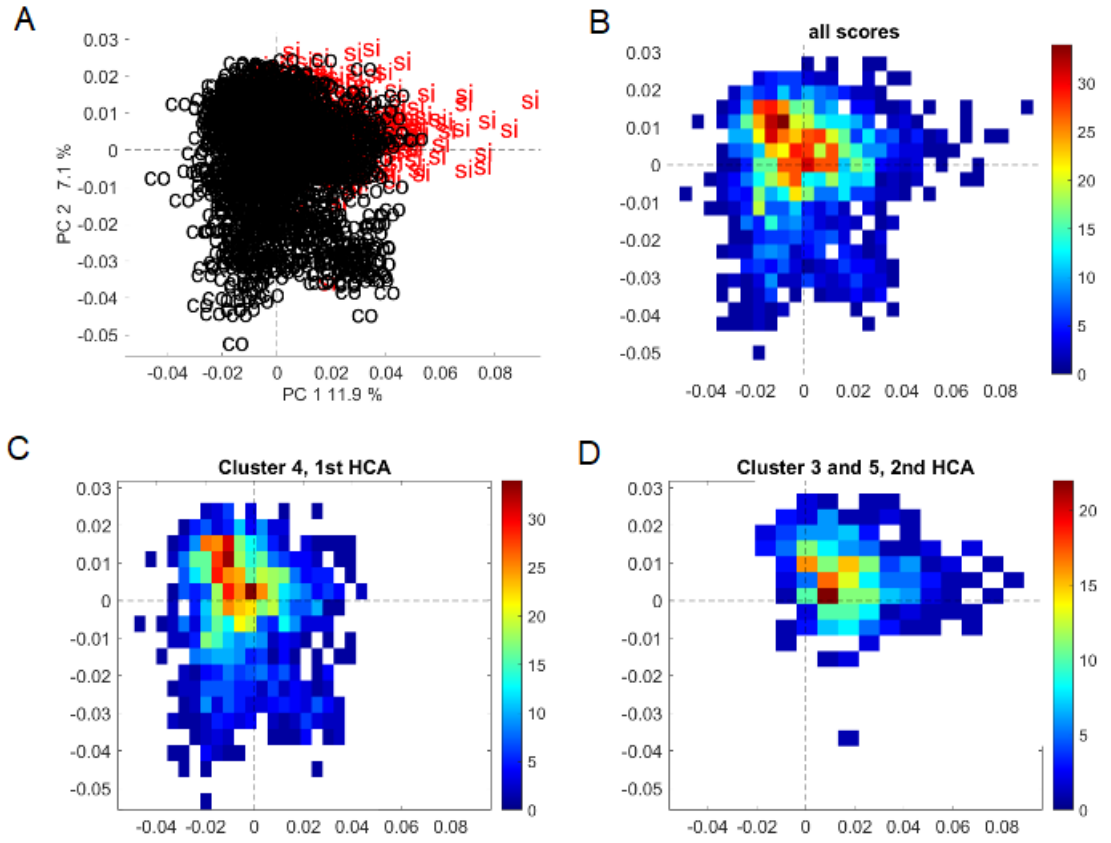

Figure 9.6: (**A**) Scores plot of the first and second PC for the 371 spectra assigned to the Si deposition spots (Cluster 3 and 5 in Figure 9.5, D) si, red) and 1541 spectra of Cluster 4 in Figure 9.5, C) , black. The Spectra were interpolated using 1.7 data points and the spectral range from 400 - 1800cm^{-1}, AsLS-corrected and vector normalized. (**B**) 2D histogram for the total amount of score values presented in **A**. (**C**) 2D histogram of score values from spectra assigned to Cluster 4 in Figure 9.5, C). (**D**) 2D histogram of score values from spectra assigned to the Si deposition spots (Cluster 3 and 5 in Figure 9.5, D).

9.2.2 Discrimination of different plant organs and growth conditions by Raman imaging data

As in the data sets in Chapter 4, 5, and 7, the variances in the data set Cucumber can be displayed in a hierarchical framework (Figure 9.7). The aim of the classification here is the changes in the chemical composition of the plant organs stems, leaves, and roots. Figure 9.7 shows the overview of the total amount of maps and spectra. The whole data set contains Raman spectra from 17 maps of stems, 23 maps of leaves, and 16 maps of roots from a total of four plants (Figure 9.7).

To reduce the total amount of 168064 single spectra, a univariate selection was applied using the intensity in the spectral region of 1550 and 1700 cm^{-1} (compare with 9.1.1). As a result, the data set is reduced to 71523 single spectra (42.6 % of total amount) consisting of 24685 single spectra from stems (47.8 % of total amount of stem spectra), 16543 single spectra from

leaves (32.2 % of total amount of leaf spectra), and 30355 single spectra from roots (46.5 % of total amount of root spectra).

A PCA of the 71523 single spectra was executed and the scores plot and loadings are presented in Figure 9.8. Technically, when creating the scores plot, the values are plotted class-by-class, beginning with beginning with the score values of the leaves (Figure 9.8, A, black triangles) followed by the score values of the stems (Figure 9.8, A, red circles) and the score values for the roots (Figure 9.8, A, blue diamonds). Therefore, the score values of the leaves (Figure 9.8, A, black triangles) and the majority of the score values of the stems (Figure 9.8, A, red circles) are not visible, due to the resolution of the scores plot and the overlapping score values of the roots (Figure 9.8, A, blue diamonds).

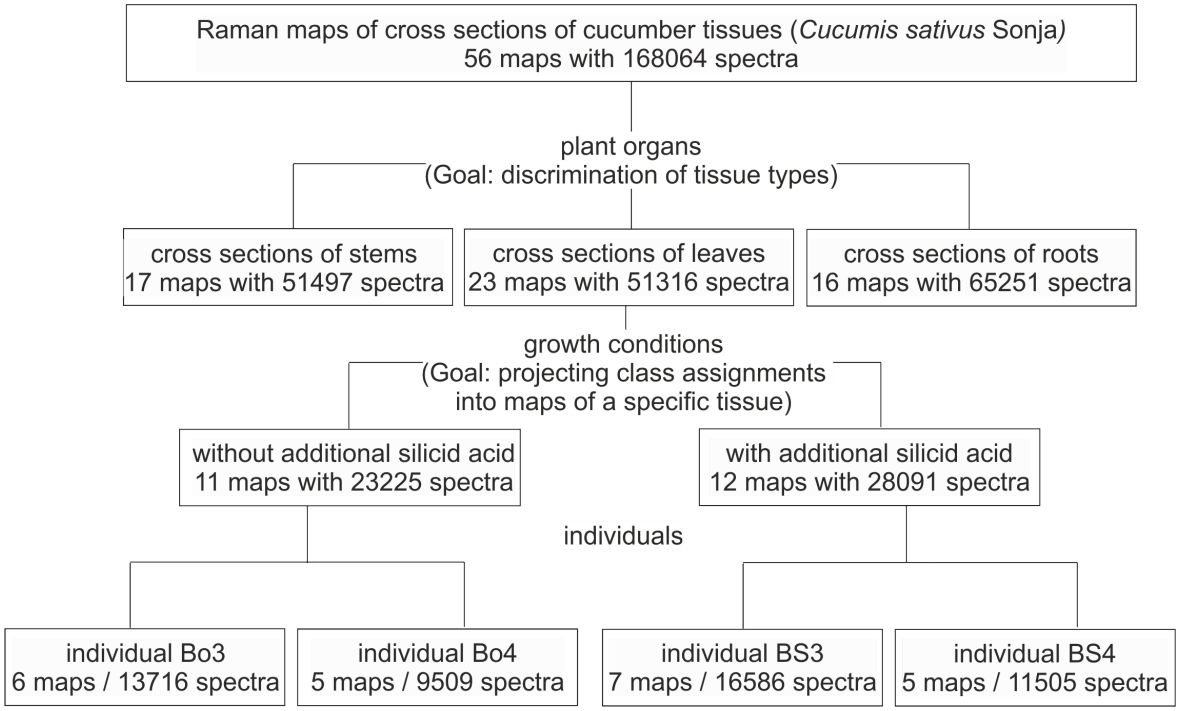

Figure 9.7: [Schematic representation of the complete data set Cucumber before any selection.] Schematic representation of the complete data set Cucumber before any selection. The variance in the group of samples is structured in a hierarchical framework, comprising three plant organs from plants grown under two conditions and from four individuals. The main focus in this classification experiment is the discrimination of spectra from different plant organs and the idea of applying an additional classification parameter in the individual maps without claiming biological significance of separation regarding growth conditions or individuals.

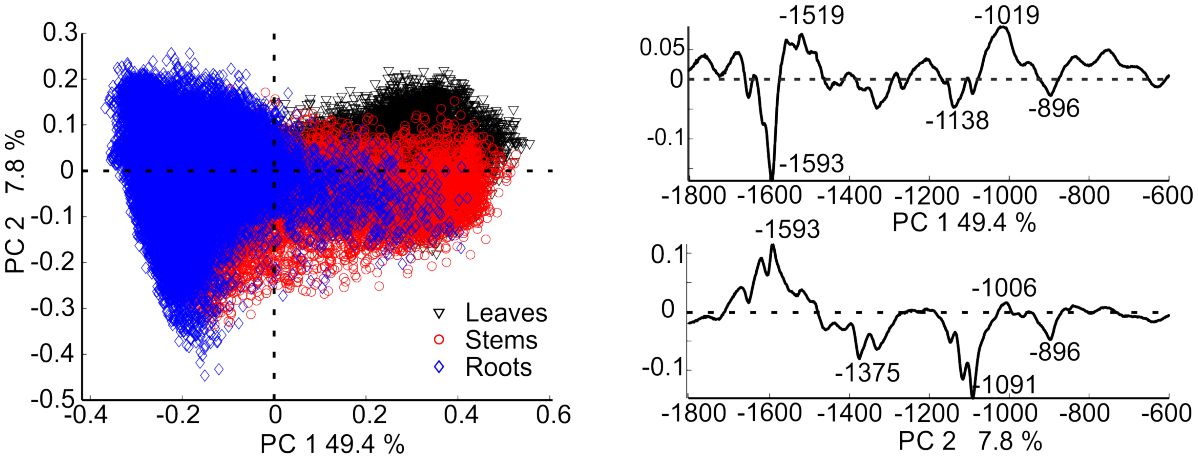

Figure 9.8: (**Left**) Scores plot and **(right)** loadings of the first and second PC for the 71523 spectra from the different plant organs, leaves (black triangles), stems (red circles), and roots (blue diamonds). The spectra were interpolated using 1.8 data points and the spectral range from 600 - 1800cm^{-1}, AsLS-corrected and vector normalized.

The according 2D histogram (Figure 9.9 (A)) indicates the occurrence of two maxima, the global maximum with negative score values regarding the first PC and a maximum with positive values regarding the first PC. Therefore it can be concluded, that a large group of spectra can be discriminated from another group of spectra from this data set.
As discussed above, this discussion about all score values in one histogram is not sufficient, since the sources of variance is unknown. Figure 9.9, B, C, and D present the 2D histograms of score values from leaf spectra (Figure 9.9, B), stem spectra (Figure 9.9, C), or root spectra (Figure 9.9, D), respectively.
The three histograms indicate a different distribution including different maxima for each group of score values. It should be emphasized that most of the score values for the leaf spectra (Figure 9.9, B) have positive values regarding PC 1, whereas most of the score values of the root spectra (Figure 9.9, D) have negative values regarding PC 1. Since the bins are colored blue, which correspond with a low number of score values, for the negative score values in the case of the leaf spectra and for the positive score values of the root spectra, it can be concluded, that the amount of overlapping score values is negligible small. The position of the maximum of the distribution of the score values of leaf spectra (Figure 9.9, B) is similar to the maximum of the distribution of all score values. (Figure 9.9, A). Likewise, the position of the maximum of the score values of the root spectra (Figure 9.9, D) can be compared with the position of the global maximum of the distribution of all score values. (Figure 9.9, A).
The 2D histrogram of the score values from stem spectra is displayed in Figure 9.9, C. The score values are distributed over the whole plot, since the score values of the stem spectra show positive and negative values regarding the first PC. The distribution show two maxima, one for the positive values and one for negative values regarding the first PC. In this example, the main variances in the spectra can be assigned to the dissimilarities in leaf and root spectra, since the maxima of their distribution can be separated regarding the first PC.

The loadings in Figure 9.8 B show an influence of the bands at 1138 (n-alkanes of suberin[152,218]) and 1593 cm^{-1} (aryl ring stretching of lignin[6,239]), as well as 896 cm^{-1} (HCC and HCO bending in cellulose[239]), 1019, and 1519 cm^{-1}. The bands can be assigned as lignin and cellulose,[6] which indicates, that the root and leaf spectra show differences in their cell wall composition.

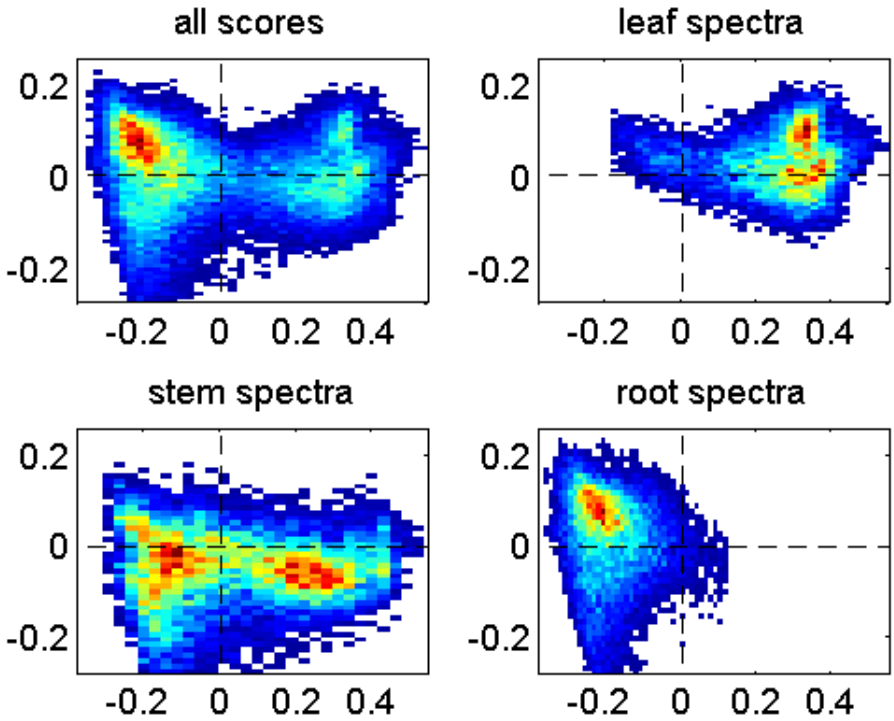

Figure 9.9: (A) 2D histogram of the score plot presented in Figure 9.8 A. **(B)** 2D-histgrams of the separated plotted score values of leaf spectra, **(C)** stem spectra, and **(D)** root spectra based on Figure 9.8 A.

2D histograms can help to interpret large data sets including spectra from several Raman maps combined, in cases where the investigated source of variance is high such as different plant organs. Here, the analysis of leaf spectra, obtained from plants grown under different conditions will be discussed (Figure 9.7). The subset contains 23 Raman maps of leaf cross sections, with 11 maps and 7524 selected spectra (32.4 % of the total amount of leaf spectra from plants growing in soil without additional silicic acid) from plants growing in soil without additional silicic acid and 12 maps and 9019 selected spectra (32.1 % of the total amount of leaf spectra from plants growing in soil with additional silicic acid) from plants growing in soil with additional silicic acid.

The score plot of the first and second PC, as well as the 2D histograms are presented in Figure 9.10. No differentiation between the two growth conditions regarding the first PC, which explained 30.5 % of the total variance, can be seen. The 2D histograms of the separated distributions of the two groups of score values show their global maxima at different positions

regarding PC 2. The global maximum of the distribution of score values from the leaf spectra of plants grown without silicic acid (Figure 9.10, C) is positive regarding PC 2, whereas the global maximum of the distribution of score values from the leaf spectra of plants grown with silicic acid (Figure 9.10, D) is negative regarding PC 2.

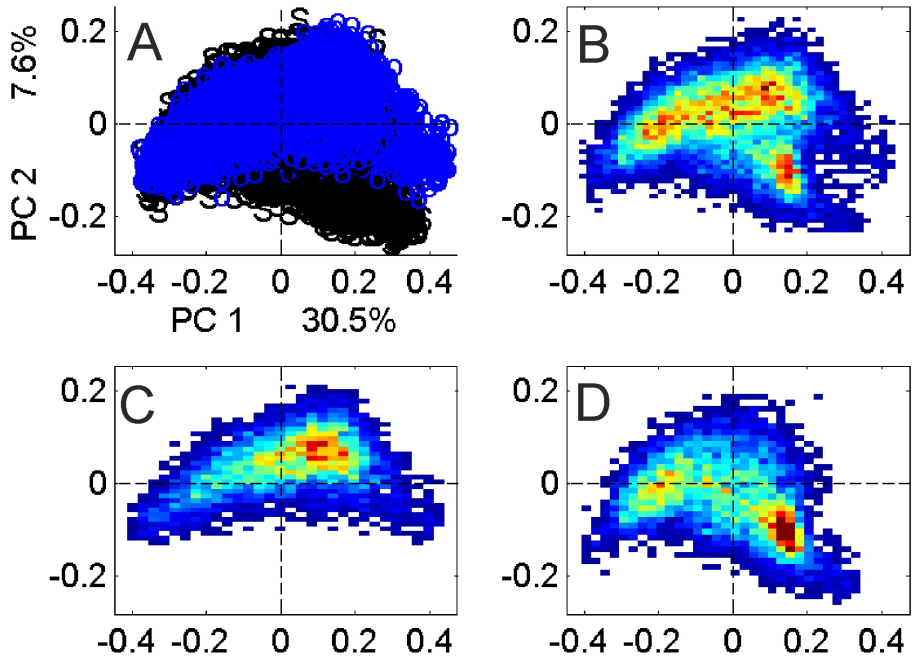

Figure 9.10: (A) Score plot of the first and second PC for the 16543 selected spectra from leaves of plants growing under different conditions with additional silicic acid in the soil, black and without additional silicic acid in the soil, blue. The spectra were interpolated using 1.8 data points and the spectral range from 600 - 1800cm^{-1}, AsLS-corrected and vector normalized. **(B)** 2D histogram for the total amount of score values presented in A. **(C)** 2D histogram of score values of leaf spectra from plants without additional silicic acid. **(D)** 2D histogram of score values of leaf spectra from plants with additional silicic acid.

To discriminate between the plants grown under different conditions the interpretation of the scores plot using PC1 and PC2 is limited. Since each of the Raman maps contributes their local variances to the total amount of variances the visualization method of the results needs another presentation of the PCA results based on each Raman map respectively.

For classification of spectra, HCA was executed using the subset of score values from PC2 to PC4. The score values were assigned to a Cluster 1 or a Cluster 2 and rearranged as the images for each of the 23 Raman maps.
In Table 9.1 an overview of the amount of selected spectra and assigned spectra is presented for each Raman map. The majority of the spectra from the Raman maps of plants grown with silicic acid are assigned as Cluster 1 (Table 9.1, first section), whereas most of the spectra from

the Raman maps of plants grown without silicic acid are clustered as Cluster 2 (Table 9.1, first section).

The clustering results presented in Table 9.1 can be visualized in each Raman map using the original local information of each spectrum. Figure 9.11 shows such maps for each of the leaf sections obtained after HCA using three principal components (PC 2 - PC 4). The visualization helps to identify spectra, that are classified in one of the two clusters. The spectra of the Raman images in the first row are mostly assigned to Cluster 1. In addition, three out of six images show more spectra assigned to Cluster 1 than assigned to Cluster 2. These first two rows are images from leaves, where the plants grow with additional silicic acid in the irrigation water. Below the black line in Figure 9.11, the clustering results of the Raman images from plants without additional silicic acid are presented. The majority of spectra in each Raman map are assigned to Cluster 2.

Table 9.1: Overview of the Raman maps and selected spectra and the amount of spectra that were clustered into Cluster 1 and Cluster 2 using the PCA score values of the second to fourth component. Selection was executed using the intensity of the spectral region 1550 - 1700cm^{-1} and a threshold of 0.25. HCA was conducted using Eudlidean distances and *Ward's* algorithm.

sample	total amount of spectra	selected amount of spectra	amount of spectra Cluster 1	amount of spectra Cluster 2
BS3A	1984	761 (38.3 %)	608 (79.9 %)	153 (20.1 %)
BS3B	2021	660 (32.6 %)	625 (94.7 %)	35 (5.3 %)
BS3C	2548	469 (18.4 %)	392 (83.6 %)	77 (16.4 %)
BS3D	2523	989 (39.20 %)	985 (99.6 %)	4 (0.4 %)
BS3E	2484	1041 (41.9 %)	1031 (99.0 %)	10 (1.0 %)
BS3F	2376	933 (39.3 %)	865 (92.7 %)	68 (7.3 %)
BS3G	2650	708 (26.7 %)	519 (73.2 %)	189 (26.8 %)
BS4A	2080	605 (29.1 %)	236 (39.0 %)	369 (61.0 %)
BS4B	2418	768 (31.7 %)	389 (50.6 %)	379 (49.4 %)
BS4C	4092	1107 (27.0 %)	332(30.0 %)	775 (70.0 %)
BS4D	759	242 (31.9 %)	41 (16.9 %)	201 (83.1 %)
BS4E	2156	736 (34.1 %)	438 (59.5 %)	298 (40.5 %)
Bo3A	3366	1192 (35.4 %)	142 (11.9 %)	1050 (88.1 %)
Bo3B	1827	624 (34.1 %)	145 (23.2 %)	479 (76.8 %)
Bo3C	3135	1036 (33.0%)	131 (12.6 %)	905 (87.4 %)
Bo3D	1596	494 (31.1 %)	71 (17.7 %)	423 (82.3 %)
Bo3E	2112	690 (32.7 %)	22 (3.9 %)	668 (96.6 %)
Bo3F	1680	568 (33.8 %)	167 (29.4 %)	401 (70.6 %)
Bo4A	1519	511 (33.6 %)	53 (11.4 %)	458 (89.6 %)
Bo4B	2430	751 (30.9 %)	273 (36.3 %)	478 (63.7 %)
Bo4C	1715	512 (29.9 %)	161 (32.4 %)	351 (68.6 %)
Bo4D	1836	639 (34.8 %)	74 (11.6 %)	565 (88.4 %)
Bo4E	2009	505 (25.1 %)	257 (50.9 %)	248 (49.1 %)

In principle, this mapping approach can give insight into the variation with a group of spectra across many maps. Although the plants that grew with and without silicic acid suggest differences in the cell wall composition, it should be pointed out that any potential biological interpretation requires a higher number of individual plants in order to attain significance. The connection of silicification and cell wall composition, including Raman spectroscopic experiments, is a highly interesting topic that is currently addressed in other projects.[39,40]

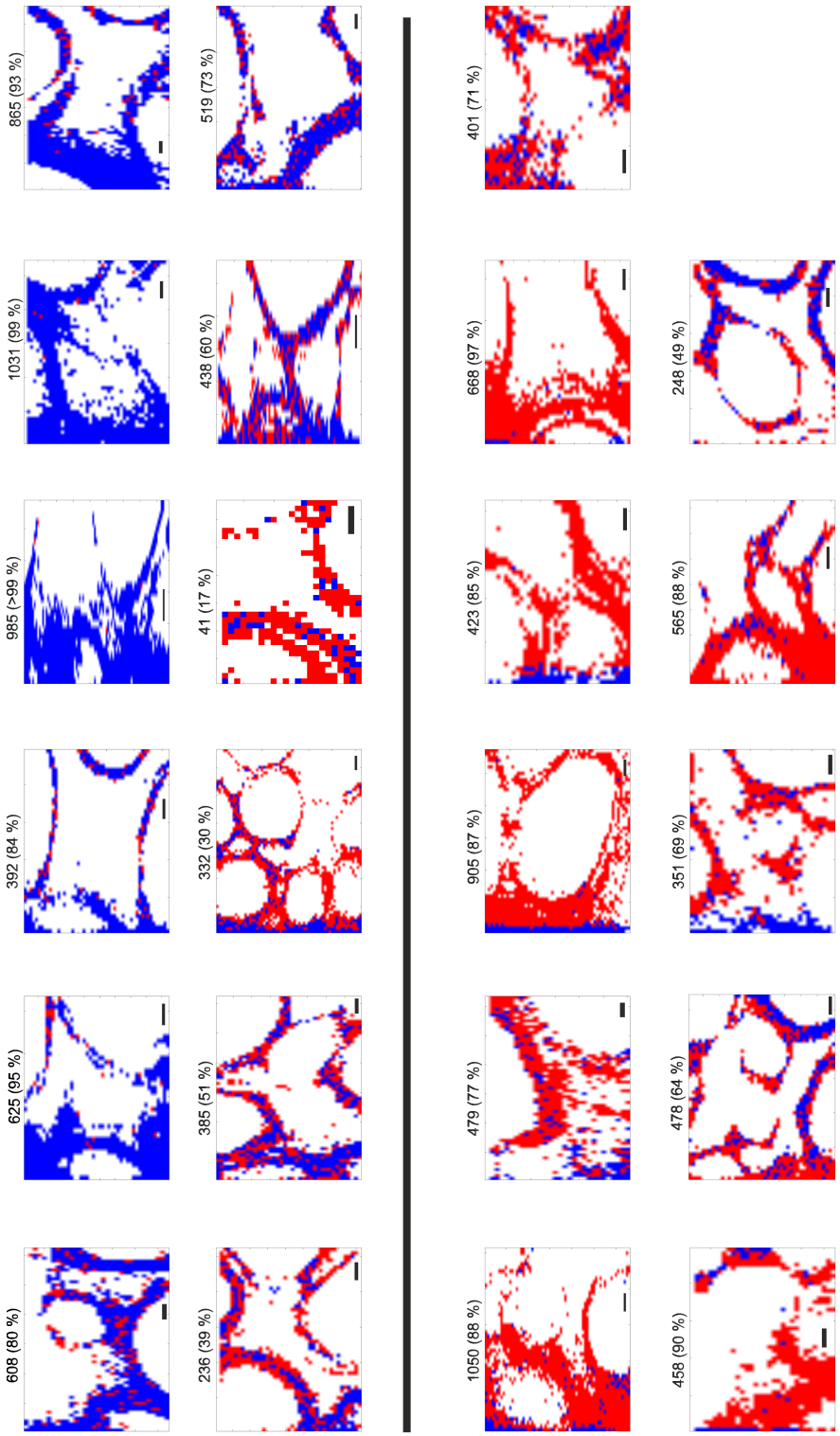

Figure 9.11: Classification results of the 23 Raman maps using the score values of PC2-PC4 and afterwards HCA. The images are colored by the two biggest clusters defined as Cluster 1 (blue) and Cluster 2 (red) (see also Table 9.1. Top, cross sections from plants with additional silicic acid. Bottom, cross sections from plants without additional silicic acid. Images are compressed for visualization. Scale bar, 10 μm.

9.2.3 Using Raman imaging data from plants in multiblock analyses

PCA and consensus principal component analysis (CPCA) are applied to a data set that comprises Raman maps from leaf cross sections of *Sorghum bicolor* wild-type and mutant (*SbLsi1*-mutant[37]). Figure 9.12 gives an overview of the data set that was used in this section. The whole data set comprises 8 individual plants and 18 Raman maps divided into nine maps of mutant plants (*SbLsi1*-mutant) from four individual plants and nine maps of wild-type plants from four individual plants of *Sorghum bicolor*. Here, the classification problem is if and how the tissue substructures of mutant plants differ from the same tissue substructures of wild-type plants. Specifically, the possibility to combine the Raman spectroscopic mapping data with additional plant data, similarly to the approach discussed in Chapter 5 is explored. The data were obtained in collaboration with the project partners Prof. Dr. Rivka Elbaum, Ingrid Liedtke and Dr. Zsuzsanna Heiner.
Following the multivariate selection approach discussed in Section 9.1.2, the data were sorted into clusters that can be assigned to certain plant tissue substructures. HCA imaging with three or five clusters for each Raman map separately was executed (compare with Figure 9.4). The clusters, were manually assigned to a Cluster Lumen or to the cell wall region divided into one cluster representing the border region (Cluster Border) and one cluster for the middle lamella (Cluster Middle) for each Raman map (Table 9.2). It should be pointed out that for some maps the cell wall region could not be divided into the Border and Middle clusters. In these cases (Table 9.2, bold) cell wall spectra are assigned to both clusters.
The data set contains a large amount of spectra, that could be analyzed using the 2D histogram visualization discussed above (cf. Figure 9.9 and Figure 9.10). In order to combine the spectroscopic information from the mapping data together with additional information from the individual plants, the spectra were averaged before data analysis.
In addition to the three spectral data sets (Border, Middle, Lumen based on the multivariate selection approach) additional plant information was measured by the cooperation partner. Energy-dispersive X-ray spectroscopy (EDX) measurement on the ashes of the seventh leaves (numbering of the leaves as proposed by Kumar *et al.*[193]), of the eight plants, were executed by Ingrid Liedtke and the relative amount of the elements carbon, calcium, magnesium, potassium, and sodium were obtained. Furthermore, plant related data is discussed separately as a data set, which includes the area of the seventh leaf of the plant, the cell wall density, the height of the plant, and the mass of the dried seventh leaf. The cell wall density was calculated by the ratio of the cell wall spectra (Border and Middle) and Lumen spectra, and the total amount of spectra.

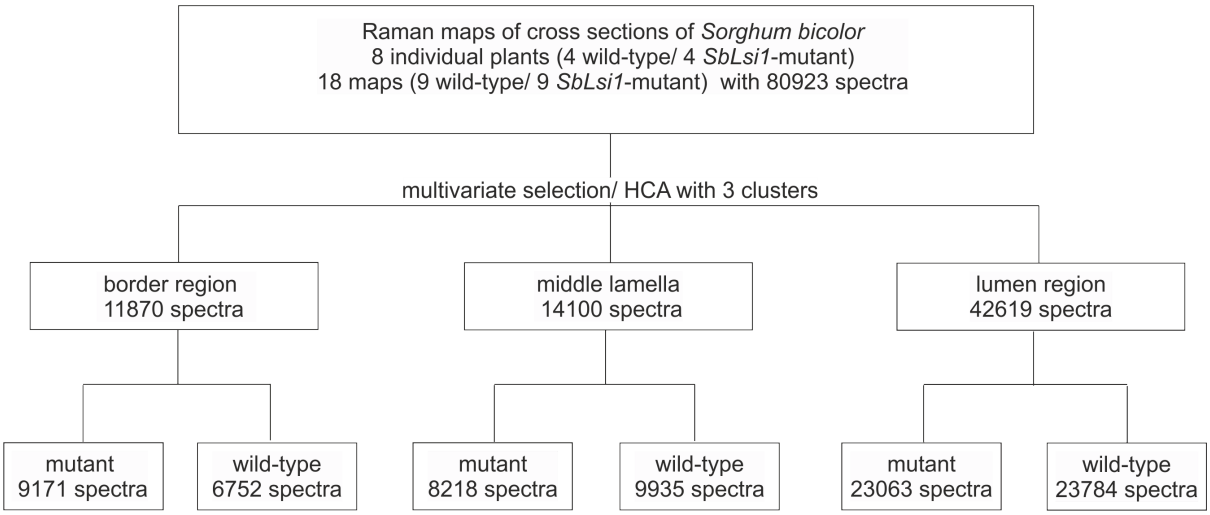

Figure 9.12: Schematic representation of the complete data set Sorghum. The data set comprises 18 Raman maps from 8 individual plants.

To compare the spectroscopic characteristics with the additional data, the spectra were averaged for each plant, which leads to eight spectra for each of the three spectral data sets, Border, Middle and Lumen.

PCA was carried out using the eight averaged spectra of each data set. The results are summarized in Figure 9.13. All loadings show a high explained variance within PC 1 and a high contribution of the noise in the PC 2 (Figure 9.13, right column). In Figure 9.13(A), scores plots and loadings of PC 1 and PC 2 are presented for the Border data set. All score values from image spectra from wild-type plants have positive values regarding PC 1 and PC 2, whereas the score values of spectra from the mutant samples are distributed over the whole scores plot. Regarding the first PC, three out of four score values for the wild-type spectra are grouping together and show a separation towards most of the mutant samples. The corresponding loadings indicate an influence of the bands at 1168, 1519, and 1598 cm^{-1}, which can be assigned to lignin (1168 and 1598 cm^{-1}) and -for unknown reasons- to carotenoids carotenoids (1519 cm^{-1}).[80] The loadings of PC 2 in Figure 9.13(A, PC 2) show features that can be assigned to cellulose (1337 cm^{-1}[237,239]) and lignin building blocks (1168, 1591, 1602, and 1627 cm^{-1}[237,239]) These bands describe the differences between the spectra of two mutants (MUB2 and MUB4, see Table 9.2) with negative values for PC 2 from the other spectra assigned as Cluster Border Figure 9.13(A).

Table 9.2: Overview of the 18 Raman maps from the data set Sorghum and the amount of spectra assigned to one of the clusters Border, Middle, and Lumen using HCA with using Euclidean distances and *Ward's* algorithm. For spectra assigned to the Clusters Border and Middle of samples MU2B1 and MU2B2, the same spectra are used in the separate classifications of the two respective substructures (in bold).

sample	total amount of spectra	spectra in Cluster Border	spectra in Cluster Middle	spectra in Cluster Lumen
MU1B1	3008	591 (19.6 %)	282 (9.4 %)	2135 (71.0 %)
MU1B2	2448	262 (10.7 %)	179 (7.3 %)	2007 (82.0 %)
MU1B3	4935	666 (13.5 %)	626 (12.7 %)	3643 (73.8 %)
MU2B1	**3920**	**2280** (58.2 %)	**2280** (58.2 %)	1640 (41.8 %)
MU2B2	**4361**	**1773** (40.6 %)	**1773** (40.6 %)	2588 (59.3 %)
MU4B1	4422	92 (2.1 %)	865 (19.6 %)	3844(86.9 %)
MU4B2	4134	1095 (26.5 %)	622 (15.1 %)	2417 (58.5 %)
MU5B1	4941	894 (18.1 %)	922 (18.7 %)	3125 (63.2 %)
MU5B2	4230	1518 (35.9 %)	1048 (24.8 %)	1664 (39.3 %)
WT1B1	2795	155 (5.5 %)	1304 (11.9 %)	1336 (88.1 %)
WT1B2	4473	1626 (36.3%)	1067 (23.8 %)	1780 (39.8 %)
WT1B3	4560	202 (4.4%)	1579 (34.6 %)	2779 (60.9 %)
WT2B1	5040	455 (9.0 %)	249 (4.9 %)	4336 (86.0 %)
WT2B2	3880	758 (19.5 %)	320 (8.2 %)	2802 (72.2 %)
WT3B1	5876	1349 (23.0 %)	893 (15.2 %)	3634 (61.8 %)
WT3B2	4100	308 (7.5 %)	3187 (77.7 %)	605 (14.8 %)
WT4B1	4560	742 (16.3 %)	938 (20.6 %)	2880 (63.1 %)
WT4B2	5187	1157 (22.3 %)	398 (7.7 %)	3632 (70.0 %)

Out of the eight spectra from Cluster Middle, five spectra (two spectra from mutant plants and three spectra from wild-type plants) show high similarity to each other, as indicated by positive score values for the first and the second PC. As in the case of the spectra of the Cluster Border the main differences are described by the signals at 1168, 1519, and 1598 cm^{-1}. In fact, the loading of PC 1 looks very similar to the loading of PC 2 in Figure 9.13 (A, right). The two spectra of the mutant plants that have negative values regarding PC 1 are from the same Raman image as the two spectra from the Cluster Border of the mutant spectra with negative score values (Figure 9.13 (A, left)). Nevertheless, the loadings of PC 2 of the middle spectra (Figure 9.13 (B, right)) show indeed extrema at the same positions as the bands above (1166, 1591, and 1604 cm^{-1}, that can be assigned to lignin), but also more noise in the spectral range 600 and 1000 cm^{-1}, which causes one of the wild-type spectra to be an outlier (Figure 9.13 (B, left, wt1, red)).

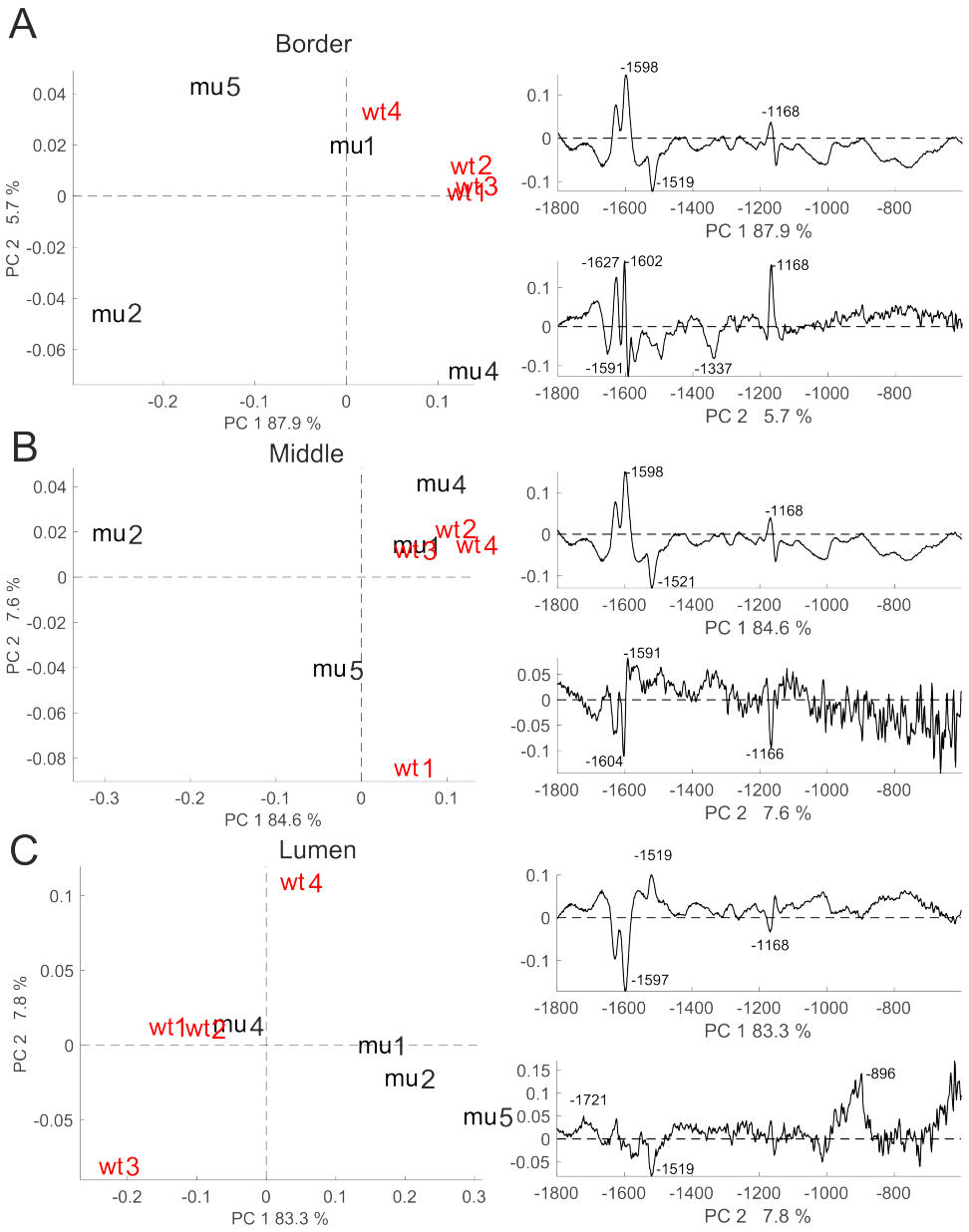

Figure 9.13: Score plot of the first and second PC for the 3 different extracted spectra regions **(A)** Border, **(B)** Middle, and **(C)** Lumen from Raman imaging data. 15923 spectra of the border regions,18153 spectra from the middle lamella, and 46847 spectra from the lumen region were interpolated and pre-processed using AsLS-correction and vector normalization before averaging to eight spectra for each data set, corresponding to the eight individual plants.

The PCA of the eight spectra from the lumen region is presented in (Figure 9.13, c). PC 1 discriminate between three out of four spectra from wild-type plants from three out of four mutant plants. Also in the case of the lumen spectra, the loadings of the first component are similar as in the case of the border and middle lamella area (1168, 1519 and 1597 cm^{-1}, that can be assigned to lignin).[237] The separation of the two specific spectra from the leaves of the *SbLsi*-mutant (compare with Figure 9.13 (A and B, left, mu2 and mu5, black) is not visible here.

From the PCA results on the averaged spectra, it can be concluded that the chemical composition within the leaves differs with respect to the individual plant. The heterogeneity in the composition of the cell wall from the *SbLsi*-mutant is higher compared to the cell wall spectra from wild-type plants (Figure 9.13 (A and B, left)). Due to the small size of the data set, the explained variance for PC 1 is above 80 % and PC 2 shows already the variation in the noise for all three regions of the cross sections. Furthermore, the loadings of PC 1 show the same features regardless of the assignment to a specific cluster, whereas the loadings of PC 2 differ with respect to the affiliated cluster of the spectra. The dividing of each Raman map, and subsequently the data set, into three clusters, enhances the probability to discriminate between individual plants.
The additional data sets obtained by EDX and morphological investigations of the plant (area of the leaf, height, cell wall density, and dry mass) were analyzed using PCA. The results for both are presented in Figure 9.14. Both scores plots indicate a separation with respect to the first PC between mutant and wild-type plants based on their respective data. In the case of the five EDX values, three out of four score values corresponding to mutant plants have negative values regarding PC 1 (Figure 9.14). The additional plant data, which includes the variables are sufficient to separate between mutant and wild-type plants.

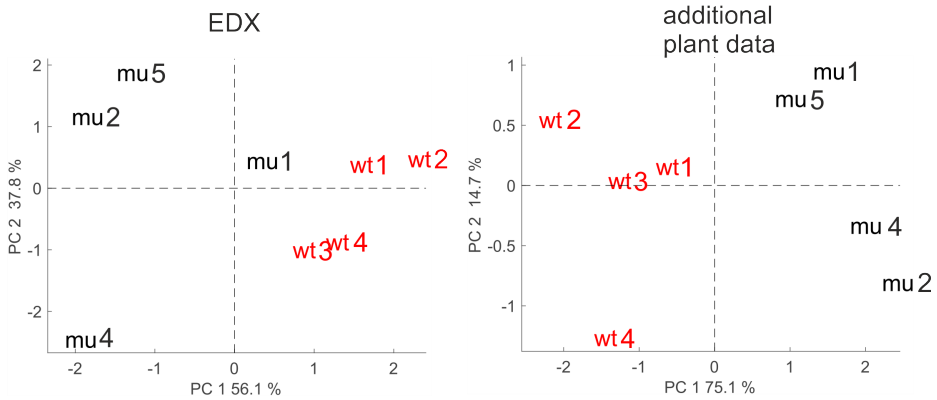

Figure 9.14: Score plots of the first and second PC for the two data sets **(Left)** EDX and **(right)** additional plant data (area of the leaf, height, cell wall density, and dry mass). Data were auto-scaled before applying PCA.

In conclusion, the separation of wild-type and *SbLsi*-mutant spectra is challenging to interpret using the averaged spectra of each individual plant. Due to a small amount of spectra, the individual variance becomes large and causes separation of individual spectra from a particular group rather than a separation based on mutant and wild-type alone. The two data sets EDX and additional plant data can help to discriminate between the different plants. A phenotyping of plants based on morphological data is a common procedure to investigate e.g. environmental influences on plants.[244,245]

A combination of data obtained from spectroscopic and non-spectroscopic data can be

conducted using CPCA. The global and block score plots are shown in Figure 9.15. The global scores indicate a separation of the data obtained from *SbLsi*-mutant and wild-type plants. The scores are separated using the first CPC, due to positive values for the wild-type plants and negative values for most of the data from mutant plants (Figure 9.15, A). CPC 1 explains 60.64 % of the variance, indicating a slightly higher influence of the three spectral data blocks compared to the discrete data blocks EDX and additional plant data. Since all three spectral blocks show similar patterns in their individual PCA (cf. Figure 9.13) high influence of the blocks on the global pattern was expected.
All block scores indicate a separation of the data from *SbLsi*-mutant and wild-type plants (Figure 9.15, B-F). For the two data blocks, EDX and additional plant data, all wild-type plant data have positive values regarding the first CPC, while the data from mutant plants have negative values (Figure 9.15, B and C). In the cases of the three spectral data blocks, the two different plant types can be separated using the first and in addition the second CPC.

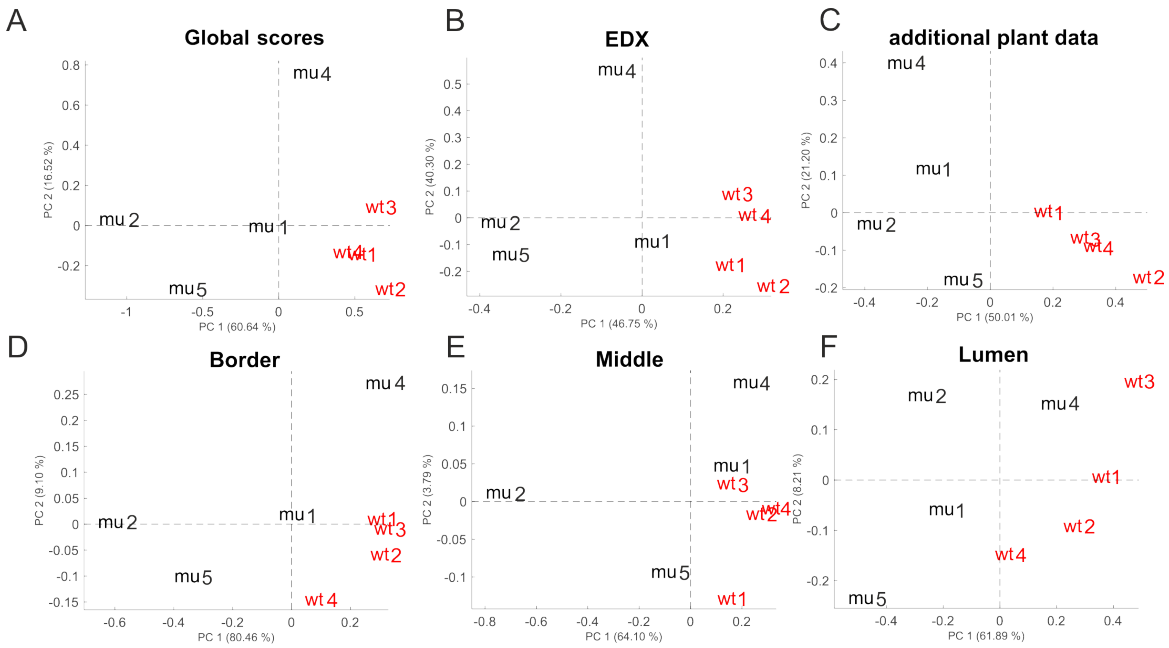

Figure 9.15: Score plots of the first and second CPC for **(A)** the global scores **(B-F)** and the blocks scores for CPCA using five data blocks consisting the two district data blocks **(B)** EDX, and **(C)** additional plant data, as well as the three data blocks from different Raman imaging regions of the cross section: **(D)** Border, **(E)** Middle, and **(F)** Lumen. 15923 spectra of the border regions as well as 18153 spectra from the middle lamella and 46847 spectra from the lumen region were interpolated and pre-processed using AsLS-correction and vector normalization before averaging to 8 spectra for each data set.

The differences between *SbLsi*-mutant and wild-type plants can be described using a correlation loadings plot. In Figure 9.16 the score values, the variables from EDX and additional plant data, as well as the extrema for the three spectral regions are displayed. The EDX variables (calcium, carbon, magnesium, potassium, and sodium are colored in magenta. The

elemental contribution of K, Ca, and Mg are mostly separating one individual *SbLsi*-mutant plant (Figure 9.16, mu4) from the rest and are therefore highly negatively correlated to each other. Moreover, this one individual mutant plant is highly correlated to the Raman band at 1336 cm^{-1} from the border region and middle lamella. This band can be assigned to carbohydrates.[239] In principle, a mineral deficiency in a plant can lead to a different carbohydrate content and/or composition,[246] and therefore could be a reason for this correlation.
The wild-type plants can be characterized by the additional plant data regarding their height, dry mass, and leaf size. These three variables are also positively correlated to the content of Mg and Ca. The three spectral data blocks from the different compartments show a positive correlation to the wild-type score values with their Raman bands that can be assigned to lignin building blocks (1168, 1593, 1602, and 1625 cm^{-1}).[239] Most of these bands are present in all three spectral data blocks.
The mutant plant leaves are positively correlated to a higher content of carbon and sodium. In addition, the mutant plants are highly correlated with a high cell wall density, which is here the ratio between spectra, that were assigned to the cell wall and the whole amount of spectra for each map. A higher cell wall density is therefore correlated with the content of carbon within the ashes. It can be concluded, that the cell wall in mutant plants are thicker, indicated by a higher amount of cell wall spectra and more carbon in the ashes. The Raman band at 1134/1139 cm^{-1} of the border region and middle lamella is correlated with high carbon content and high cell wall density and can be assigned to suberin.[152,218] In addition several Raman bands that occur in all three spectral data blocks are highly correlated to the mutant plants. The Raman bands at 761, 769, 1519, 1670, and 1685 cm^{-1} are in addition to the mutant plant score values highly correlated to the content of sodium in the ashes. The bands can be assigned as building blocks of lignin. Therefore, it can be suggested that the cell walls of the mutant plants differ in their composition of different lignin building blocks. In conclusion, plants from *SbLsi*-mutant and wild-type differ in their appearance, resulting in bigger leaves and higher plants for wild-type plants, as well in their cell wall composition with thicker cell walls in the case of mutant plants. CPCA enables a comprehensive analysis of the differences in the cell wall composition.

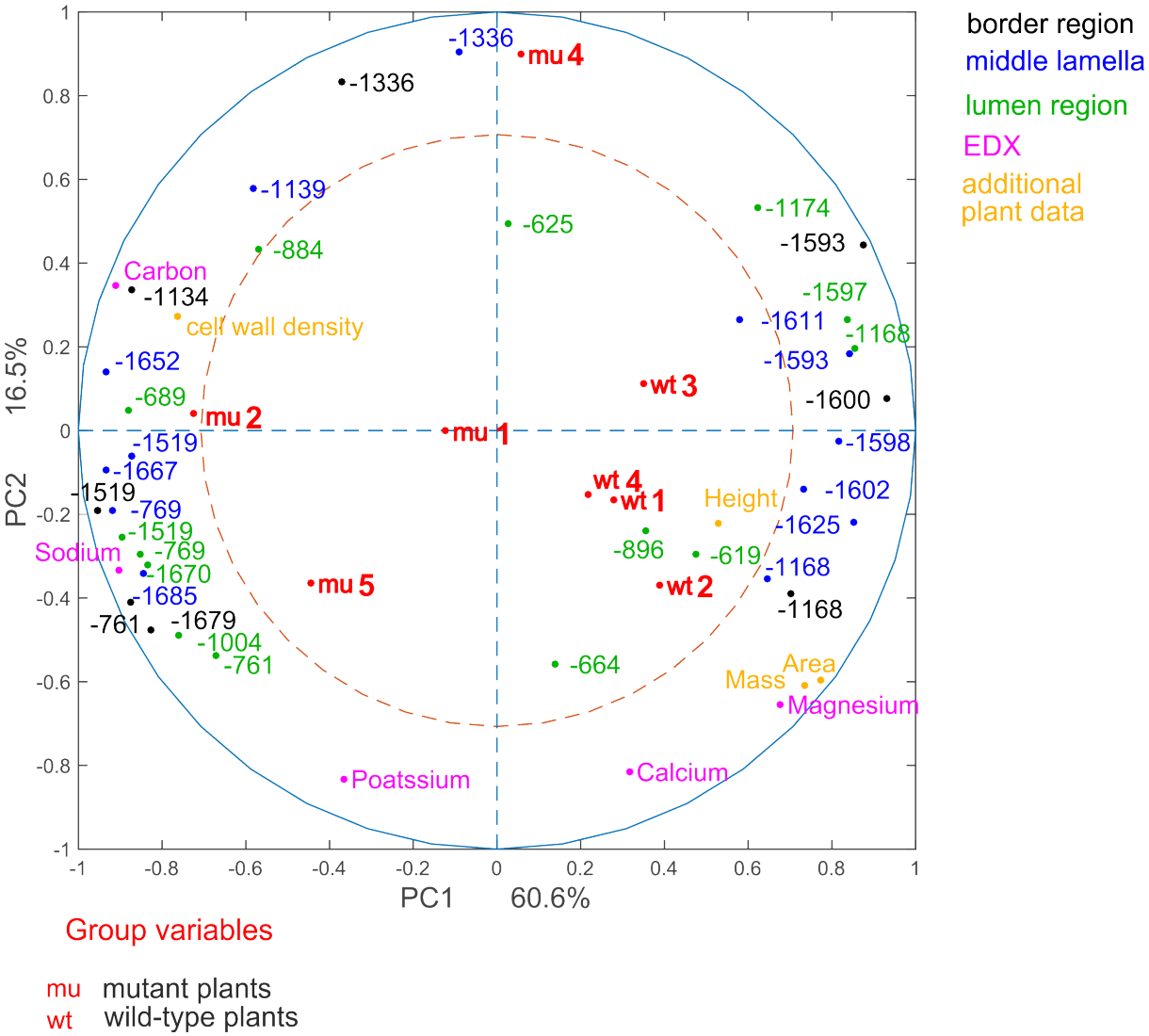

Figure 9.16: Correlation loadings plot for the first and second CPC. Displayed are the global scores of wildtype and mutant plants (red), as well as the loadings of the blocks of the border region, middle lamella and lumen region, EDX and addional plant data. For clarity only extrema of the loadings are shown for the spectroscopic data.

In this chapter, the data treatment and analysis of different Raman imaging data sets from plant tissues were presented. Two different selection approaches were discussed that can help to extract relevant spectra with respect to a biological question. In an univariate approach, the Raman spectra in images are roughly extracted according to a defined threshold. In terms of fast and automatic extraction, it is a beneficial approach. A multivariate selection should be preferred, when the spectra/imaging regions cannot be defined as relevant or non-relevant in order to discuss the classification question. A clustering approach was suggested, which can group certain spectra and mark them, in addition to the extraction.

As another important result, an approach was proposed for the recombination of class assignments in cluster analyses of many maps, leading to the projection of classification results into the original microspectroscopic map.

Selection and subsequent data analysis were demonstrated using the data sets from three different examples, namely i) the exploratory data analysis of one Raman map from a root tissue from *Sorghum bicolor* to identify and cluster important region from the map, ii) the assessment of variation within a hierarchical framework of cross sections from *Cucumis sativus* Sonja comprising different plant organs and growth conditions, and iii) the combination of spectroscopic data of plant tissue substructures with additional information on the plants for the characterization of different types (*SbLsi*-mutant and wild-type) of *Sorghum bicolor*.

10 Summary and Outlook

In this thesis, spectrometric and spectroscopic data of plant materials, in particular pollen and cross sections of plant organs, were utilized and analyzed. The focus was pt on the assessment of variances within the data regarding a specific biological question with the overall goal of combining different types of spectra for an improved utilization in plant science and bioanalysis, such as the discrimination between closely-related grass species or variations in pollen chemistry within the same species, or mapping of a specific biochemical composition in biomineralizing tissues. In some of these examples, the biological experiment had been designed in such a way that the variation within the data sets can be organized in hierarchical frameworks, where high variances in spectra are often masking smaller ones from sub-grouped levels.

As an example of such a well-designed hierarchically structured data set, in Chapter 4, pollen mass spectra obtained by matrix-assisted laser desorption/ionization time of flight mass spectrometry (MALDI-TOF MS), were assessed by principal component analysis (PCA) and partial least square discriminant analysis (PLS-DA). By applying statistical tools, here the Kruskal-Wallis[199] H-test and the Bartlett test[199,200] the PCA results were explored in an optimized way by calculating one p-value and one d-value that can describe the distribution of the score values and the dimensionality, respectively.
To assess variation within the data, classification by PLS-DA was attained. The classification model confirms the high species-specificity, as well as the high population-specificity of the pollen mass spectra. The variation between different growth conditions, which would be a third level of variances is the hierarchical framework cannot be evaluated using the whole data set. Nevertheless, MALDI-TOF MS enables the monitoring of the phenotypic plasticity/rigidity of the parental plants indicated by higher/lower success rates and by PCA when analyzing each population separately. In addition, it was shown, that a phenotypic rigidity of the pollen samples corresponds to a high discrimination in the genotypes of the respective population.
MALDI-TOF MS is well established for the identification of fungi[84,94–97] or micro-organisms,[87–89,247] but in spite of first successful efforts[1,2,56,107] still needs to be further developed for pollen identification. A rapid, reliable monitoring of pollen populations and even phenotypic plasticity would be a powerful tool in crop science, plant biology, agricultural science, and climate

studies. Since MALDI-TOF MS can enable the detection of climatic or location-related influences, statements about environmental pollution, climate change, or the crop yield of plants may be made. Nevertheless, the species- and population-specific pattern are not fully elucidated yet. MS/MS experiments could reveal more information about the nature of the peaks, which was already successfully applied on peaks in the lower mass range by Krause *et al.*[1]

Different spectroscopic and spectrometric methods, namely: FTIR spectroscopy on bulk samples, Raman spectroscopy on single pollen grains, surface enhanced Raman scattering (SERS) on water-soluble extracts, and MALDI TOF MS on acid pollen extracts were evaluated with respect of the discrimination between three pollen populations and four different growth conditions for the grass species *Poa alpina* in Chapter 5. Applied separately, all of these methods are well known to enable discrimination of different pollen species.[1–3,5,13] Here, the extent to which these methods, alone and in combination, can be used to evaluate subspecies-variation was studied. For this data set, the four methods differ in their performance. While MALDI-TOF MS enables the assessment of variation of populations, variation induced by different growth condition can be studied by FTIR, Raman, and SERS. Moreover, the same data sets can have a different influence on the distinction between different growth conditions in the three populations. This implies that, due to the different fraction of the pollen chemistry that is represented by each data set (or analyzed by each of the methods), the biochemical effect of the growth conditions on pollen chemistry can vary for different populations. This would be in agreement with variation in phenotypic plasticity between the populations, in particular regarding different metabolic and molecular pathways for environmental adaptation as discussed also in Chapter 4 for MALDI mass spectra and discussed by Zimmermann *et al.* in the case of FTIR spectroscopy.[15]
The combination of the methods using Consensus principal component analysis (CPCA)reveals comprehensive insights into the pollen composition. As an example, for one population with a higher phenotypic plasticity, a higher starch and lipid content of the pollen grains was conducted.
Comprehensive studies of pollen grains are important in many fields, including agriculture and crop science. Climate change may cause cause losses in crops and even the extinction of several plant species, therefore it is valuable to predict and characterize the phenotypic plasticity of plants based on their pollen composition. Particularly, a combination of spectroscopic/spectrometric data and further pollen characteristics, such as pollen quantity, viability, germination rate, and tube length would lead to great insights into quality. In this way, crops may be optimized and plants can be rescued from extinctions.
A multimodal study, such as the one presented here, can be conducted on other plant tissues as well. This is also indicated by the results of a CPCA that uses Raman mapping data together with other data from plants for classification (Chapter 9, see also below). In future experi-

ments, CPCA can be applied on multimodal data sets from other plant tissues as well, such as phytoliths or cell wall compartments. A novel multimodal study on phytoliths pointed out the great potential applying several different spectroscopic methods to such samples.[39] CPCA would be beneficial in the investigation of the formation of phytoliths.

Chapter 6 addresses some challenges in the utilization of Raman mapping data of single pollen grains. This concerned both the optimization of experiments in order to avoid the generation of unwanted variance, as well as data pre-treatment strategies. In experiments, the use of versatile sample substrates and the influence of unwanted pollen germination were studied. The Raman experiments were carried out on a sample set with a complex structured variation including wild-type and two different mutants (*SbLsi1* and *bmr*), as well as stressed and not stressed samples of *Sorghum bicolor*. Fixation of the pollen grains could have many advantages, since Raman spectroscopy is usually non-destructive and non-invasive. The pollen sample could be measured with different methods such as energy-dispersive X-ray spectroscopy (EDX), where fixation on carbon tape is required or MALDI-TOF MS.[55,107] In terms of MALDI-TOF MS, the fixation on carbon tape leads to good classification results, even with only a few pollen grains.[55] With increasing excitation time, the fluorescence background in the Raman mapping spectra from pollen was found to be decreasing. The digital reduction of the fluorescence background is necessary,[173,248,249] and was achieved here by asymmetric least square baseline correction, vector normalization, and averaging the spectra for each pollen grain. The wild-type/mutant variation could be assessed using the Raman mapping data after this pre-processing. Raman microscopy. CPCA of Raman spectroscopy in combination with MALDI-TOF MS leads to further insights into the differences in the chemical composition of pollen from mutant and wild-type.
Such a rather small sample set containing here 25 samples from 25 plants, respectively, is not sufficient to build a reliable model using classification methods such as PLS-DA or artificial neural networks. Particularly, non-related variation such as different growth seasons, pollen collection times and measurement time also contribute to the variance in the data set. A large greenhouse experiment with more consistent conditions would lead to more reliable models. Nevertheless, the problem of too small/biased data sets is commonly known to data scientists in bio-analytics and diagnostics. Therefore, new approaches are tested, where parts of the data set are,e.g, simulated from existing data sets[?] or no traditional training set is needed in the classification.[228]

The utilization of FTIR microspectroscopic data from single pollen grains was one of the aims of the experiments discussed in Chapter 7. FTIR measurements of single pollen grains lead to Mie scattering artifacts in pollen spectra that can be reduced by embedding the pollen grains in paraffin. Here, an adapted procedure described by Zimmermann et. al.[14] was applied in order to discriminate between pollen spectra from five closely related grass species.

As a result, additional bands of the paraffin can be found in the FTIR spectra of pollen. Four data pre-processing approaches were presented, namely i) without any considerations of a possible influence of the paraffin signals, ii) omitting of the most affected spectral range, iii) separating paraffin contribution from the pollen spectra, and iv) minimization of the paraffin constituent by applying a complex EMSC model. The latter three approaches deal with diminishing of the paraffin contribution. It leads to the successful discrimination of the pollen species *Anthoxanthum odoratum, Bromus inermis, Hordeum bulbosum, Lolium perenne*, and *Poa alpina* using PLS-DA, artificial neural networks, and Rand dom forest
Several studies deal with digital paraffin removing using chemometric approaches, e.g., independent component analysis[250,251] or EMSC.[252] The presented approaches can also be applied to other embedded tissues, such as plant cross sections. With a suitable data preprocessing a chemical de-paraffination is not required.
A combination of FTIR spectra from single pollen grains with MALDI mass spectra of pollen extracts using CPCA reveals a better classification of the pollen species discussed here. In future experiments, the combination of a well-designed and characterized data set of FTIR data and MALDI mass spectra could lead to more insight into the species-specific peak patterns observed in MALDI-TOF MS. Here, the CPCA results confirm a high correlation between several species-specifc peaks in MALDI and molecular information obtained by FTIR microspectroscopy.
A low reproducibility of the FTIR measurements from the embedded pollen grains was obtained by testing a PLS-DA model on a data set from an identical sample set measured at a different time. This needs to be considered when building FTIR databases in the future. Here, a PLS-DA model was trained using one data set and additional data including more species that were measured at a different time. Compared to a model based on MALDI mass spectra of the same sample set, the success rates are rather low and, therefore, not sufficient for automatic pollen identification. To improve the classification, FTIR spectra from several experiments and technical replicates would be required. Particularly, other classification methods could be applied to such an extended FTIR data set, e.g., more advanced model geometries, such as hierarchically ordered taxonomic classification by partial least squares (Hot-PLS)[229]

While FTIR and Raman spectroscopy could identify pollen on the single pollen grain level, in MALDI-TOF MS experiments, extracts from several single pollen grains are mixed, which leads to more challenges regarding the identification of pollen. In Chapter 8, some of these challenges, e.g, ion suppression effects[230,253] and masking of species-specific peak patterns[57] are addressed. Chemometric approaches are presented and compared that enable the identification of specific species in a pollen mixture. PLS-DA, ANN, and random forest perform with similar classification outcome visualized by classification images of mixture spectra. In addition, a model using a small database with 16 different pollen species was built and

applied on the mixtures. The results imply that the pollen species can only be identified within the mixture, if the species can also be described by the applied model. Pollen species with highly species-specific pattern are dominating the identification results.
Specifically, mixture spectra can be decomposed using NMF. The relative contributions of one component can indicate the distribution of the pollen extract on the target. In this example, decomposing a MALDI MS image of a pollen mixture of three different pollen species, two of the species can be identified by comparing the decomposed components and reference spectra from the same species.
In order to develop an automatic pollen identification system using MALDI-TOF MS, a database can be built similar to these from bacteria. Automatic identification of pollen species could be presented with a corresponding confidence interval. Here, the analysis was focused on the application of PLS-DA, ANN, RF, and NMF. Also HCA and PCA show promising results in the utilization of mass spectra from pollen mixture.[55,57] In addition other methods may lead to good or complementary results, e.g., Parallel factor analysis (PARAFAC) or k-means clustering.[236,254]
Also, first high resolution MALDI MS experiments were conducted on very small amounts of pollen grains, where overlapping of the extract from different pollen species was minimized.[55]

Chapter 9 provides strategies for multivariate data analysis of Raman imaging data when particularly spectra from many maps are involved. Usually large data sets are obtained by Raman imaging experiments for the exploration of plant tissues' substructures. Here, the analysis was conducted on three different data sets from projects studying the silicification of plant tissues.[7,40,152] One of them aims here is a comprehensive characterization of *SbLsi1* mutant and wild-type *Sorghum bicolor* plants using the spectral information from different tissuesâ€™ substructure and additional plant-related data

The investigation of smaller effects, e.g., the influence of growth condition on the composition of the cell walls, often do not require spectral information from the whole Raman map. Here, two approaches were presented in order to select the spectra in a semi-automatic way. The variance within the mapping data was reduced in a way that a classification of spectra from many Raman maps only includes a subset of spectra, e.g., from a specific tissue substructure. The first approach follows the idea of a chemical image where one spectrum is represented with one value of the key parameter, e.g., intensity of a specific band. Based on this key parameter a threshold can be defined and all spectra with a value above/below the threshold can be extracted. The approach is suitable for fast automatic extraction of spectra from maps, where the differences between the required spectra and the remaining spectra are high and a very accurate selection of spectra is not needed to solve the classification problem.
The second approach discusses a selection of spectra using multivariate methods, shown for HCA. The spectra of each map can be assigned to clusters and the results can be visualized as

an image. According to specific knowledge about the investigated tissue, the cluster could be assigned to a specific substructure. The targeted extraction of these clusters leads to data sets in which variances in, e.g., spectra from the same substructures of many maps or spectra from different substructures from the same map can be evaluated.
PCA of large data sets results in visualization problems. Here, the analysis was interpreted using 2D histograms of the score values. After exploration of the large data set comprising the mapping data of many samples, using PCA, the classification results, here obtained by a combination of PCA and HCA, can be visualized in each single map using the original spatial information. Using additional plant data and spectroscopic information, variation caused by both the environmental and the genetic differences were studied by CPCA. The presented approaches can be applied to data from several projects where the assessment of variances from many maps is needed.

The discussion of the data in this thesis indicates the possibilities that open up when the utilization of spectrometric and spectroscopic data from plant tissues is optimized regarding a specific analytical question. Using spectral data sets of very different plant and tissue origin, from three vibrational spectroscopic and one mass spectrometric approach, as well as from different spectral sampling regimes, the great importance of an optimized data pre-treatment was shown. This includes pre-processing, elimination of unwanted spectral contributions (e.g., from embedding medium or substrate), and the targeted extraction of relevant information. When assessing the variance within the data sets by mulitvariate tools, the combination of the chemical data with additional plant-related information shows great potential in the classification and characterization of subspecies variation. Specifically, a multiblock analysis that includes data from microspectroscopic and mapping experiments was demonstrated in studies on the influence of sub-species variation in the different plant tissues. Overall, this work underpins the potential of a comprehensive, many spectroscopies based chemical analysis of plant tissues, with applications in several fields.

Bibliography

[1] B. Krause, S. Seifert, U. Panne, J. Kneipp, and S. M. Weidner. Matrix-assisted laser desorption/ionization mass spectrometric investigation of pollen and their classification by multivariate statistics. *Rapid Commun Mass Spectrom*, 26(9):1032–8, 2012.

[2] S. Seifert, S. M. Weidner, U. Panne, and J. Kneipp. Taxonomic relationships of pollens from matrix-assisted laser desorption/ionization time-of-flight mass spectrometry data using multivariate statistics. *Rapid Commun Mass Spectrom*, 29(12):1145–54, 2015.

[3] Stephan Seifert, Virginia Merk, and Janina Kneipp. Identification of aqueous pollen extracts using surface enhanced raman scattering (sers) and pattern recognition methods. *Journal of Biophotonics*, 9(1-2):181–189, 2016.

[4] C. S. Pappas, P. A. Tarantilis, P. C. Harizanis, and M. G. Polissiou. New method for pollen identification by ft-ir spectroscopy. *Appl Spectrosc*, 57(1):23–7, 2003.

[5] B. Zimmermann. Characterization of pollen by vibrational spectroscopy. *Appl Spectrosc*, 64(12):1364–73, 2010.

[6] Notburga Gierlinger and Manfred Schwanninger. Chemical imaging of poplar wood cell walls by confocal raman microscopy. *Plant Physiology*, 140(4):1246–1254, 2006.

[7] Ingrid Zeise, Zsuzsanna Heiner, Sabine Holz, Maike Joester, Carmen Büttner, and Janina Kneipp. Raman imaging of plant cell walls in sections of cucumis sativus. *Plants*, 7(1):7, 2018.

[8] 1st international plant spectroscopy conference, 2017.

[9] Adele C. M. Julier, Phillip E. Jardine, Angela L. Coe, William D. Gosling, Barry H. Lomax, and Wesley T. Fraser. Chemotaxonomy as a tool for interpreting the cryptic diversity of poaceae pollen. *Review of Palaeobotany and Palynology*, 235:140–147, 2016.

[10] M. Bagcioglu, B. Zimmermann, and A. Kohler. A multiscale vibrational spectroscopic approach for identification and biochemical characterization of pollen. *Plos One*, 10(9):e0137899, 2015.

[11] M. L. Laucks, G. Roll, G. Schweiger, and E. J. Davis. Physical and chemical (raman) characterization of bioaerosolsâ€"pollen. *Journal of Aerosol Science*, 31(3):307–319, 1999.

[12] N. Ivleva, R. Niessner, and U. Panne. Characterization and discrimination of pollen by raman microscopy. *Analytical and Bioanalytical Chemistry*, 381(1):261–267, 2005.

[13] F. Schulte, J. Lingott, U. Panne, and J. Kneipp. Chemical characterization and classification of pollen. *Analytical Chemistry*, 80(24):9551–9556, 2008.

[14] B. Zimmermann and A. Kohler. Infrared spectroscopy of pollen identifies plant species and genus as well as environmental conditions. *Plos One*, 9(4), 2014.

[15] B. Zimmermann, M. Bagcioglu, V. Tafinstseva, A. Kohler, M. Ohlson, and S. Fjellheim. A high-throughput ftir spectroscopy approach to assess adaptive variation in the chemical composition of pollen. *Ecol Evol*, 7(24):10839–10849, 2017.

[16] Maike Joester, Stephan Seifert, Franziska Emmerling, and Janina Kneipp. Physiological influence of silica on germinating pollen as shown by raman spectroscopy. *Journal of Biophotonics*, 10(4):542–552, 2017.

[17] J. Depciuch, I. Kasprzyk, E. Roga, and M. Parlinska-Wojtan. Analysis of morphological and molecular composition changes in allergenic artemisia vulgaris l. pollen under traffic pollution using sem and ftir spectroscopy. *Environ Sci Pollut Res Int*, 23(22):23203–23214, 2016.

[18] J. Depciuch, I. Kasprzyk, O. Sadik, and M. Parlinska-Wojtan. Ftir analysis of molecular composition changes in hazel pollen from unpolluted and urbanized areas. *Aerobiologia (Bologna)*, 33(1):1–12, 2017.

[19] V. Joseph, F. Schulte, H. Rooch, I. Feldmann, I. Dorfel, W. Osterle, U. Panne, and J. Kneipp. Surface-enhanced raman scattering with silver nanostructures generated in situ in a sporopollenin biopolymer matrix. *Chem Commun (Camb)*, 47(11):3236–8, 2011.

[20] J. Rozema, R. A. Broekman, P. Blokker, B. B. Meijkamp, N. de Bakker, J. van de Staaij, A. van Beem, F. Ariese, and S. M. Kars. Uv-b absorbance and uv-b absorbing compounds (para-coumaric acid) in pollen and sporopollenin: the perspective to track historic uv-b levels. *J Photochem Photobiol B*, 62(1-2):108–17, 2001.

[21] Peter Blokker, Peter Boelen, Rob Broekman, and Jelte Rozema. The occurrence of p-coumaric acid and ferulic acid in fossil plant materials and their use as uv-proxy. *Plant Ecology*, 182(1):197, 2006.

[22] Phillip E. Jardine, Feargus A. J. Abernethy, Barry H. Lomax, William D. Gosling, and Wesley T. Fraser. Shedding light on sporopollenin chemistry, with reference to uv reconstructions. *Review of Palaeobotany and Palynology*, 238:1–6, 2017.

[23] F. S. Li, P. Phyo, J. Jacobowitz, M. Hong, and J. K. Weng. The molecular structure of plant sporopollenin. *Nat Plants*, 5(1):41–46, 2019.

[24] P. Piffanelli, J. H. E. Ross, and D. J. Murphy. Biogenesis and function of the lipidic structures of pollen grains. *Sexual Plant Reproduction*, 11(2):65–80, 1998.

[25] S. Wang, D. Wang, Q. Wu, K. Gao, Z. Wang, and Z. Wu. 3d imaging of a rice pollen grain using transmission x-ray microscopy. *J Synchrotron Radiat*, 22(4):1091–5, 2015.

[26] T. D. Macfarlane and L. Watson. The classification of poaceae subfamily pooideae. *Taxon*, 31(2):178–203, 1982.

[27] Robert J. Soreng, Paul M. Peterson, Konstantin Romaschenko, Gerrit Davidse, Fernando O. Zuloaga, Emmet J. Judziewicz, Tarciso S. Filgueiras, Jerrold I. Davis, and Osvaldo Morrone. A worldwide phylogenetic classification of the poaceae (gramineae). *Journal of Systematics and Evolution*, 53(2):117–137, 2015.

[28] Maduraimuthu Djanaguiraman, P. V. Vara Prasad, Marimuthu Murugan, Ramasamy Perumal, and Umesh K. Reddy. Physiological differences among sorghum (sorghum bicolor l. moench) genotypes under high temperature stress. *Environmental and Experimental Botany*, 100:43–54, 2014.

[29] Joseph H. Williams and Susan J. Mazer. Pollenâ€"tiny and ephemeral but not forgotten: New ideas on their ecology and evolution. *American Journal of Botany*, 103(3):365–374, 2016.

[30] Y. Jiang, R. Lahlali, C. Karunakaran, S. Kumar, A. R. Davis, and R. A. Bueckert. Seed set, pollen morphology and pollen surface composition response to heat stress in field pea. *Plant Cell Environ*, 38(11):2387–97, 2015.

[31] F. Schulte, J. Mader, L. W. Kroh, U. Panne, and J. Kneipp. Characterization of pollen carotenoids with in situ and high-performance thin-layer chromatography supported resonant raman spectroscopy. *Analytical Chemistry*, 81(20):8426–8433, 2009.

[32] B. Zimmermann, M. Bagcioglu, C. Sandt, and A. Kohler. Vibrational microspectroscopy enables chemical characterization of single pollen grains as well as comparative analysis of plant species based on pollen ultrastructure. *Planta*, 242(5):1237–50, 2015.

[33] N. Gierlinger and M. Schwanninger. The potential of raman microscopy and raman imaging in plant research. *Spectroscopy-an International Journal*, 21(2):69–89, 2007.

[34] C. J. G. Colares, T. C. M. Pastore, V. T. R. Coradin, J. A. A. Camargos, A. C. O. Moreira, J. C. Rubim, and J. W. B. Braga. Exploratory analysis of the distribution of lignin and cellulose in woods by raman imaging and chemometrics. *Journal of the Brazilian Chemical Society*, 26(6):1297–1305, 2015.

[35] Sara Piqueras, Sophie Füchtner, Rodrigo Rocha de Oliveira, AdriÃ¡n Gomez-Sanchez, Stanislav Jelavic, Tobias Keplinger, Anna de Juan, and Lisbeth Garbrecht Thygesen. Understanding the formation of heartwood in larch using synchrotron infrared imaging combined with multivariate analysis and atomic force microscope infrared spectroscopy. *Frontiers in Plant Science*, 10(1701), 2020.

[36] Batirtze Prats-Mateu, Martin Felhofer, Anna de Juan, and Notburga Gierlinger. Multivariate unmixing approaches on raman images of plant cell walls: new insights or overinterpretation of results? *Plant Methods*, 14(1):52, 2018.

[37] Oshry Markovich, Santosh Kumar, Dikla Cohen, Sefi Addadi, Eyal Fridman, and Rivka Elbaum. Silicification in leaves of sorghum mutant with low silicon accumulation. *Silicon*, 11(5):1–7, 2015.

[38] Santosh Kumar, Milan Soukup, and Rivka Elbaum. Silicification in grasses: Variation between different cell types. *Frontiers in Plant Science*, 8(438):438, 2017.

[39] V. M. R. Zancajo, S. Diehn, N. Filiba, G. Goobes, J. Kneipp, and R. Elbaum. Spectroscopic discrimination of sorghum silica phytoliths. *Front Plant Sci*, 10(1571):1571, 2019.

[40] Nerya Zexer and Rivka Elbaum. Unique lignin modifications pattern the nucleation of silica in sorghum endodermis. *Journal of Experimental Botany*, 2020.

[41] Kimberley Gallagher, Alba Alfonso-Garcia, Jessica Sanchez, Eric Potma, and Guaciara Santos. Plant growth conditions alter phytolith carbon. *Frontiers in Plant Science*, 6(753), 2015.

[42] Gea Guerriero, Jean-Francois Hausman, and Sylvain Legay. Silicon and the plant extracellular matrix. *Frontiers in Plant Science*, 7(463), 2016.

[43] Daniel J. Cosgrove. Growth of the plant cell wall. *Nature Reviews Molecular Cell Biology*, 6(11):850–861, 2005.

[44] Diao She, Feng Xu, ZengChao Geng, RunCang Sun, Gwynn Lloyd Jones, and Mark S. Baird. Physicochemical characterization of extracted lignin from sweet sorghum stem. *Industrial Crops and Products*, 32(1):21–28, 2010.

[45] Monika Szymanska-Chargot, Piotr M. Pieczywek, Monika ChyliÅ„ska, and Artur Zdunek. Hyperspectral image analysis of raman maps of plant cell walls for blind spectra characterization by nonnegative matrix factorization algorithm. *Chemometrics and Intelligent Laboratory Systems*, 151:136–145, 2016.

[46] M. Szymanska-Chargot, M. Chylinska, P. M. Pieczywek, P. Rosch, M. Schmitt, J. Popp, and A. Zdunek. Raman imaging of changes in the polysaccharides distribution in the cell wall during apple fruit development and senescence. *Planta*, 243(4):935–45, 2016.

[47] Monika Chylinska, Monika SzymaÅ„ska-Chargot, Beata Kruk, and Artur Zdunek. Study on dietary fibre by fourier transform-infrared spectroscopy and chemometric methods. *Food Chemistry*, 196:114–122, 2016.

[48] N. Gierlinger, T. Keplinger, and M. Harrington. Imaging of plant cell walls by confocal raman microscopy. *Nature Protocols*, 7(9):1694–1708, 2012.

[49] Notburga Gierlinger, Lanny Sapei, and Oskar Paris. Insights into the chemical composition of equisetum hyemale by high resolution raman imaging. *Planta*, 227(5):969–80, 2008.

[50] Umesh P. Agarwal. Raman imaging to investigate ultrastructure and composition of plant cell walls: distribution of lignin and cellulose in black spruce wood (picea mariana). *Planta*, 224(5):1141–1153, 2006.

[51] Umesh P. Agarwal, James D. McSweeny, and Sally A. Ralph. Ftâ€“raman investigation of milled-wood lignins: Softwood, hardwood, and chemically modified black spruce lignins. *Journal of Wood Chemistry and Technology*, 31(4):324–344, 2011.

[52] N. Perisic, N. K. Afseth, R. Ofstad, B. Narum, and A. Kohler. Characterizing salt substitution in beef meat processing by vibrational spectroscopy and sensory analysis. *Meat Sci*, 95(3):576–85, 2013.

[53] N. Perisic, N. K. Afseth, R. Ofstad, S. Hassani, and A. Kohler. Characterising protein, salt and water interactions with combined vibrational spectroscopic techniques. *Food Chem*, 138(1):679–86, 2013.

[54] B. Zimmermann, V. Tafintseva, M. Bagcioglu, M. Hoegh Berdahl, and A. Kohler. Analysis of allergenic pollen by ftir microspectroscopy. *Anal Chem*, 88(1):803–11, 2016.

[55] F. Lauer, S. Diehn, S. Seifert, J. Kneipp, V. Sauerland, C. Barahona, and S. Weidner. Multivariate analysis of maldi imaging mass spectrometry data of mixtures of single pollen grains. *J Am Soc Mass Spectrom*, 29(11):2237–2247, 2018.

[56] S. Weidner, R. D. Schultze, and B. Enthaler. Matrix-assisted laser desorption/ionization imaging mass spectrometry of pollen grains and their mixtures. *Rapid Commun Mass Spectrom*, 27(8):896–903, 2013.

[57] Franziska Lauer. *Massenspektrometrische Untersuchungen einzelner Pollenkörner*. Thesis, 2019.

[58] Joe H. Ward. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, 1963.

[59] Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

[60] H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6):417–441, 1933.

[61] Herman Wold. *11 - Path Models with Latent Variables: The NIPALS Approach**NIPALS = Nonlinear Iterative PArtial Least Squares*, pages 307–357. Academic Press, 1975.

[62] S Wold, S Hellberg, T Lundstedt, M Sjostrom, and H Wold. Proc. symp. on pls model building: Theory and application. In *Frankfurt am Main*.

[63] Johan A. Westerhuis, Theodora Kourti, and John F. MacGregor. Analysis of multiblock and hierarchical pca and pls models. *Journal of Chemometrics*, 12(5):301–321, 1998.

[64] Sahar Hassani, Harald Martens, El Mostafa Qannari, Mohamed Hanafi, Grethe Iren Borge, and Achim Kohler. Analysis of -omics data: Graphical interpretation- and validation tools in multi-block methods. *Chemometrics and Intelligent Laboratory Systems*, 104(1):140–153, 2010.

[65] Sahar Hassani, Mohamed Hanafi, El Mostafa Qannari, and Achim Kohler. Deflation strategies for multi-block principal component analysis revisited. *Chemometrics and Intelligent Laboratory Systems*, 120:154–168, 2013.

[66] H. Wold. Estimation of principal components and related models by iterative least squares. *Multivariate analysis*, pages 391–420, 1966.

[67] Herman Wold. Causal flows with latent variables: Partings of the ways in the light of nipals modelling. *European Economic Review*, 5(1):67–86, 1974.

[68] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–91, 1999.

[69] Karthik Devarajan. Nonnegative matrix factorization: An analytical and interpretive tool in computational biology. *PLOS Computational Biology*, 4(7):e1000029, 2008.

[70] J. Leuschner, M. Schmidt, P. Fernsel, D. Lachmund, T. Boskamp, and P. Maass. Supervised non-negative matrix factorization methods for maldi imaging applications. *Bioinformatics*, 35(11):1940–1947, 2019.

[71] Y. Gut, M. Boiret, L. Bultel, T. Renaud, A. Chetouani, A. Hafiane, Y. M. Ginot, and R. Jennane. Application of chemometric algorithms to maldi mass spectrometry imaging of pharmaceutical tablets. *J Pharm Biomed Anal*, 105:91–100, 2015.

[72] Robert Luce, Peter Hildebrandt, Uwe Kuhlmann, and Jörg Liesen. Using separable nonnegative matrix factorization techniques for the analysis of time-resolved raman spectra. *Applied Spectroscopy*, 70(9):1464–1475, 2016.

[73] Sanjeev Arora, Rong Ge, Ravi Kannan, and Ankur Moitra. Computing a nonnegative matrix factorization—provably. *SIAM Journal on Computing*, 45:1582–1611, 2016.

[74] J. Zupan and J. Gasteiger. Neural networks: A new method for solving chemical problems or just a passing phase? *Analytica Chimica Acta*, 248(1):1–30, 1991.

[75] Gary B. Fogel. Computational intelligence approaches for pattern discovery in biological systems. *Briefings in Bioinformatics*, 9(4):307–316, 2008.

[76] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[77] S. B. Kotsiantis. Decision trees: a recent overview. *Artificial Intelligence Review*, 39(4):261–283, 2013.

[78] Aakash Parmar, Rakesh Katariya, and Vatsal Patel. *A Review on Random Forest: An Ensemble Classifier*, pages 758–763. International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI) 2018. Springer International Publishing, Cham, 2019.

[79] D. Gill, R. G. Kilponen, and L. Rimai. Resonance raman scattering of laser radiation by vibrational modes of carotenoid pigment molecules in intact plant tissues. *Nature*, 227(5259):743–744, 1970.

[80] Hartwig Schulz and Malgorzata Baranska. Identification and quantification of valuable plant substances by ir and raman spectroscopy. *Vibrational Spectroscopy*, 43(1):13–25, 2007.

[81] N. Gierlinger. Revealing changes in molecular composition of plant cell walls on the micron-level by raman mapping and vertex component analysis (vca). *Frontiers in Plant Science*, 5:306, 2014.

[82] A. Gorzsas, H. Stenlund, P. Persson, J. Trygg, and B. Sundberg. Cell-specific chemotyping and multivariate imaging by combined ft-ir microspectroscopy and orthogonal projections to latent structures (opls) analysis reveals the chemical landscape of secondary xylem. *Plant J*, 66(5):903–14, 2011.

[83] Miao Liang, Peng Zhang, Xi Shu, Changgeng Liu, and Jinian Shu. Characterization of pollen by maldi-tof lipid profiling. *International Journal of Mass Spectrometry*, 334:13–18, 2013.

[84] D. F. O. Rocha, C. M. S. Cunha, K. R. A. Belaz, F. N. Dos Santos, R. H. Hinz, A. Pereira, E. Wicket, L. M. Andrade, C. A. O. Nascimento, A. Visconti, and M. N. Eberlin. Lipid and protein fingerprinting for fusarium oxysporum f. sp. cubense strain-level classification. *Anal Bioanal Chem*, 409(29):6803–6812, 2017.

[85] M. Welker. Proteomics for routine identification of microorganisms. *Proteomics*, 11(15):3143–53, 2011.

[86] 3rd Yates, J. R. Mass spectrometry and the age of the proteome. *J Mass Spectrom*, 33(1):1–19, 1998.

[87] P. Lasch, W. Beyer, H. Nattermann, M. Stammler, E. Siegbrecht, R. Grunow, and D. Naumann. Identification of bacillus anthracis by using matrix-assisted laser desorption ionization-time of flight mass spectrometry and artificial neural networks. *Appl Environ Microbiol*, 75(22):7229–42, 2009.

[88] P. Lasch, H. Nattermann, M. Erhard, M. Stammler, R. Grunow, N. Bannert, B. Appel, and D. Naumann. Maldi-tof mass spectrometry compatible inactivation method for highly pathogenic microbial cells and spores. *Anal Chem*, 80(6):2026–34, 2008.

[89] E. Nagy, T. Maier, E. Urban, G. Terhes, M. Kostrzewa, and Escmid Study Group on Antimicrobial Resistance in Anaerobic Bacteria. Species identification of clinical isolates of bacteroides by matrix-assisted laser-desorption/ionization time-of-flight mass spectrometry. *Clin Microbiol Infect*, 15(8):796–802, 2009.

[90] S. Sauer, A. Freiwald, T. Maier, M. Kube, R. Reinhardt, M. Kostrzewa, and K. Geider. Classification and identification of bacteria by mass spectrometry and computational analysis. *PLoS One*, 3(7):e2843, 2008.

[91] P. A. Demirev and C. Fenselau. Mass spectrometry for rapid characterization of microorganisms. *Annu Rev Anal Chem (Palo Alto Calif)*, 1:71–93, 2008.

[92] G. Marklein, M. Josten, U. Klanke, E. Muller, R. Horre, T. Maier, T. Wenzel, M. Kostrzewa, G. Bierbaum, A. Hoerauf, and H. G. Sahl. Matrix-assisted laser desorption ionization-time of flight mass spectrometry for fast and reliable identification of clinical yeast isolates. *J Clin Microbiol*, 47(9):2912–7, 2009.

[93] O. Bader. Maldi-tof-ms-based species identification and typing approaches in medical mycology. *Proteomics*, 13(5):788–99, 2013.

[94] Carole Cassagne, Anne-CÃ©cile Normand, Coralie L'Ollivier, StÃ©phane Ranque, and Renaud Piarroux. Performance of maldi-tof ms platforms for fungal identification. *Mycoses*, 59(11):678–690, 2016.

[95] J. Chalupova, M. Raus, M. Sedlarova, and M. Sebela. Identification of fungal microorganisms by maldi-tof mass spectrometry. *Biotechnol Adv*, 32(1):230–41, 2014.

[96] H. Ling, Z. Yuan, J. Shen, Z. Wang, and Y. Xu. Accuracy of matrix-assisted laser desorption ionization-time of flight mass spectrometry for identification of clinical pathogenic fungi: a meta-analysis. *J Clin Microbiol*, 52(7):2573–82, 2014.

[97] S. Ranque, A. C. Normand, C. Cassagne, J. B. Murat, N. Bourgeois, F. Dalle, M. Gari-Toussaint, P. Fourquet, M. Hendrickx, and R. Piarroux. Maldi-tof mass spectrometry identification of filamentous fungi in the clinical laboratory. *Mycoses*, 57(3):135–40, 2014.

[98] Stephanie Kaspar, Manuela Peukert, Ales Svatos, Andrea Matros, and Hans-Peter Mock. Maldi-imaging mass spectrometry â€“ an emerging technique in plant biology. *PROTEOMICS*, 11(9):1840–1850, 2011.

[99] C. Plomion, C. Lalanne, S. Claverol, H. Meddour, A. Kohler, M. B. Bogeat-Triboulot, A. Barre, G. Le Provost, H. Dumazet, D. Jacob, C. Bastien, E. Dreyer, A. de Daruvar, J. M. Guehl, J. M. Schmitter, F. Martin, and M. Bonneu. Mapping the proteome of poplar and application to the discovery of drought-stress responsive proteins. *Proteomics*, 6(24):6509–27, 2006.

[100] M. Burrell, C. Earnshaw, and M. Clench. Imaging matrix assisted laser desorption ionization mass spectrometry: a technique to map plant metabolites within tissues at high spatial resolution. *J Exp Bot*, 58(4):757–63, 2007.

[101] N. Imin, T. Kerim, B. G. Rolfe, and J. J. Weinman. Effect of early cold stress on the maturation of rice anthers. *Proteomics*, 4(7):1873–82, 2004.

[102] A. Jacobs, P. T. Larsson, and O. Dahlman. Distribution of uronic acids in xylans from various species of soft- and hardwood as determined by maldi mass spectrometry. *Biomacromolecules*, 2(3):979–90, 2001.

[103] P. Verdino. Structural characterization of pollen allergens. *Clin Rev Allergy Immunol*, 30(2):73–95, 2006.

[104] M. J. Raftery, R. G. Saldanha, C. L. Geczy, and R. K. Kumar. Mass spectrometric analysis of electrophoretically separated allergens and proteases in grass pollen diffusates. *Respir Res*, 4(10):10, 2003.

[105] S. G. Iraneta, D. M. Acosta, R. Duran, C. Apicella, U. D. Orlando, M. A. Seoane, A. Alonso, and V. G. Duschak. Maldi-tof ms analysis of labile lolium perenne major allergens in mixes. *Clin Exp Allergy*, 38(8):1391–9, 2008.

[106] L. P. Chow, L. L. Chiu, K. H. Khoo, H. J. Peng, S. Y. Yang, S. W. Huang, and S. N. Su. Purification and structural analysis of the novel glycoprotein allergen cyn d 24, a pathogenesis-related protein pr-1, from bermuda grass pollen. *FEBS J*, 272(24):6218–27, 2005.

[107] F. Lauer, S. Seifert, J. Kneipp, and S. M. Weidner. Simplifying the preparation of pollen grains for maldi-tof ms classification. *Int J Mol Sci*, 18(3):11, 2017.

[108] Richard M. Caprioli, Terry B. Farmer, and Jocelyn Gile. Molecular imaging of biological samples:â€‰ localization of peptides and proteins using maldi-tof ms. *Analytical Chemistry*, 69(23):4751–4760, 1997.

[109] Sarah Robinson, Karen Warburton, Mark Seymour, Malcolm Clench, and Jane Thomas-Oates. Localization of water-soluble carbohydrates in wheat stems using imaging matrix-assisted laser desorption ionization mass spectrometry. *New Phytologist*, 173(2):438–444, 2007.

[110] Manuela Peukert, Andrea Matros, Giuseppe Lattanzio, Stephanie Kaspar, Javier AbadÃa, and Hans-Peter Mock. Spatially resolved analysis of small molecules by matrix-assisted laser desorption/ionization mass spectrometric imaging (maldi-msi). *New Phytologist*, 193(3):806–815, 2012.

[111] Rohit Shroff, Katharina Schramm, Verena Jeschke, Peter Nemes, Akos Vertes, Jonathan Gershenzon, and AleÅ¡ SvatoÅ¡. Quantification of plant surface metabolites by matrix-assisted laser desorptionâ€“ionization mass spectrometry imaging: glucosinolates on arabidopsis thaliana leaves. *The Plant Journal*, 81(6):961–972, 2015.

[112] Gregor McCombie, Dieter Staab, Markus Stoeckli, and Richard Knochenmuss. Spatial and spectral correlations in maldi mass spectrometry images by clustering and multivariate analysis. *Analytical Chemistry*, 77(19):6118–6124, 2005.

[113] G. Gudi, A. Krahmer, I. Koudous, J. Strube, and H. Schulz. Infrared and raman spectroscopic methods for characterization of taxus baccata l.–improved taxane isolation by accelerated quality control and process surveillance. *Talanta*, 143:42–49, 2015.

[114] M. KacurÃ¡kovÃ¡. Ft-ir study of plant cell wall model compounds: pectic polysaccharides and hemicelluloses. *Carbohydrate Polymers*, 43(2):195–203, 2000.

[115] A. Largo-Gosens, M. Hernandez-Altamirano, L. Garcia-Calvo, A. Alonso-Simon, J. Alvarez, and J. L. Acebes. Fourier transform mid infrared spectroscopy applications for monitoring the structural plasticity of plant cell walls. *Front Plant Sci*, 5:303, 2014.

[116] D. M. Musingarabwi, H. H. Nieuwoudt, P. R. Young, H. A. Eyeghe-Bickong, and M. A. Vivier. A rapid qualitative and quantitative evaluation of grape berries at various stages of development using fourier-transform infrared spectroscopy and multivariate data analysis. *Food Chem*, 190:253–262, 2016.

[117] M. J. Baker, J. Trevisan, P. Bassan, R. Bhargava, H. J. Butler, K. M. Dorling, P. R. Fielden, S. W. Fogarty, N. J. Fullwood, K. A. Heys, C. Hughes, P. Lasch, P. L. Martin-Hirsch, B. Obinaju, G. D. Sockalingum, J. Sule-Suso, R. J. Strong, M. J. Walsh, B. R. Wood, P. Gardner, and F. L. Martin. Using fourier transform ir spectroscopy to analyze biological materials. *Nat Protoc*, 9(8):1771–91, 2014.

[118] R. Dell'Anna, P. Lazzeri, M. Frisanco, F. Monti, F. Malvezzi Campeggi, E. Gottardini, and M. Bersani. Pollen discrimination and classification by fourier transform infrared (ft-ir) microspectroscopy and machine learning. *Anal Bioanal Chem*, 394(5):1443–52, 2009.

[119] J. Depciuch, I. Kasprzyk, E. Drzymala, and M. Parlinska-Wojtan. Identification of birch pollen species using ftir spectroscopy. *Aerobiologia (Bologna)*, 34(4):525–538, 2018.

[120] Elena Gottardini, Stefano Rossi, Fabiana Cristofolini, and Luca Benedetti. Use of fourier transform infrared (ft-ir) spectroscopy as a tool for pollen identification. *Aerobiologia*, 23(3):211–219, 2007.

[121] A. Guedes, H. Ribeiro, M. Fernandez-Gonzalez, M. J. Aira, and I. Abreu. Pollen raman spectra database: application to the identification of airborne pollen. *Talanta*, 119:473–8, 2014.

[122] A. Woutersen, P. E. Jardine, R. G. Bogota-Angel, H. X. Zhang, D. Silvestro, A. Antonelli, E. Gogna, R. H. J. Erkens, W. D. Gosling, G. Dupont-Nivet, and C. Hoorn. A novel approach to study the morphology and chemistry of pollen in a phylogenetic context, applied to the halophytic taxon nitraria l.(nitrariaceae). *PeerJ*, 6:e5055, 2018.

[123] Murat Bagcioglu, Achim Kohler, Stephan Seifert, Janina Kneipp, Boris Zimmermann, and Sean McMahon. Monitoring of plant environment interactions by high throughput ftir spectroscopy of pollen. *Methods in Ecology and Evolution*, 8(7):870–880, 2017.

[124] Adriana Kendel and Boris Zimmermann. Chemical analysis of pollen by ft-raman and ftir spectroscopies. *Frontiers in Plant Science*, 11(352), 2020.

[125] M. Mularczyk-Oliwa, A. Bombalska, M. Kaliszewski, M. Wlodarski, K. Kopczynski, M. Kwasny, M. Szpakowska, and E. A. Trafny. Comparison of fluorescence spectroscopy and ftir in differentiation of plant pollens. *Spectrochim Acta A Mol Biomol Spectrosc*, 97:246–54, 2012.

[126] Phillip Jardine, William Gosling, Barry Lomax, Adele Julier, and Wesley Fraser. *Chemotaxonomy of domesticated grasses: a pathway to understanding the origins of agriculture*, volume 38. 2019.

[127] R. Lukacs, R. Blumel, B. Zimmerman, M. Bagcioglu, and A. Kohler. Recovery of absorbance spectra of micrometer-sized biological and inanimate particles. *Analyst*, 140(9):3273–84, 2015.

[128] P. Bassan, H. J. Byrne, F. Bonnier, J. Lee, P. Dumas, and P. Gardner. Resonant mie scattering in infrared spectroscopy of biological materials–understanding the 'dispersion artefact'. *Analyst*, 134(8):1586–93, 2009.

[129] T. Konevskikh, R. Lukacs, and A. Kohler. An improved algorithm for fast resonant mie scatter correction of infrared spectra of cells and tissues. *J Biophotonics*, 11(1), 2018.

[130] P. Bassan, H. J. Byrne, J. Lee, F. Bonnier, C. Clarke, P. Dumas, E. Gazi, M. D. Brown, N. W. Clarke, and P. Gardner. Reflection contributions to the dispersion artefact in ftir spectra of single biological cells. *Analyst*, 134(6):1171–5, 2009.

[131] H. Martens and E. Stark. Extended multiplicative signal correction and spectral interference subtraction: new preprocessing methods for near infrared spectroscopy. *J Pharm Biomed Anal*, 9(8):625–35, 1991.

[132] C. V. Raman and K. S. Krishnan. A new type of secondary radiation. *Nature*, 121(3048):501–502, 1928.

[133] G Landsberg and L Mandelstam. A novel effect of light scattering in crystals. *Naturwissenschaften*, 16(5):5, 1928.

[134] Adolf Smekal. Zur quantentheorie der dispersion. *Naturwissenschaften*, 11(43):873–875, 1923.

[135] W. E. Huang, R. I. Griffiths, I. P. Thompson, M. J. Bailey, and A. S. Whiteley. Raman microscopic analysis of single microbial cells. *Anal Chem*, 76(15):4452–8, 2004.

[136] Yu-San Huang, Takeshi Karashima, Masayuki Yamamoto, Takashi Ogura, and Hiro-o Hamaguchi. Raman spectroscopic signature of life in a living yeast cell. *Journal of Raman Spectroscopy*, 35(7):525–526, 2004.

[137] H. Wu, J. V. Volponi, A. E. Oliver, A. N. Parikh, B. A. Simmons, and S. Singh. In vivo lipidomics using single-cell raman spectroscopy. *Proc Natl Acad Sci U S A*, 108(9):3809–14, 2011.

[138] K. Maquelin, C. Kirschner, L. P. Choo-Smith, N. A. Ngo-Thi, T. van Vreeswijk, M. Stammler, H. P. Endtz, H. A. Bruining, D. Naumann, and G. J. Puppels. Prospective study of the performance of vibrational spectroscopies for rapid identification of bacterial and fungal pathogens recovered from blood cultures. *J Clin Microbiol*, 41(1):324–9, 2003.

[139] K. Maquelin, C. Kirschner, L. P. Choo-Smith, N. van den Braak, H. P. Endtz, D. Naumann, and G. J. Puppels. Identification of medically relevant microorganisms by vibrational spectroscopy. *J Microbiol Methods*, 51(3):255–71, 2002.

[140] A. Oust, T. Moretro, K. Naterstad, G. D. Sockalingum, I. Adt, M. Manfait, and A. Kohler. Fourier transform infrared and raman spectroscopy for characterization of listeria monocytogenes strains. *Appl Environ Microbiol*, 72(1):228–32, 2006.

[141] U. P. Agarwal. 1064 nm ft-raman spectroscopy for investigations of plant cell walls and other biomass materials. *Front Plant Sci*, 5:490, 2014.

[142] N. Altangerel, G. O. Ariunbold, C. Gorman, M. H. Alkahtani, E. J. Borrego, D. Bohlmeyer, P. Hemmer, M. V. Kolomiets, J. S. Yuan, and M. O. Scully. In vivo diagnostics of early abiotic plant stress response via raman spectroscopy. *Proc Natl Acad Sci U S A*, 114(13):3393–3396, 2017.

[143] Yu Cao, Deyan Shen, Yonglay Lu, and Yong Huang. A raman-scattering study on the net orientation of biomacromolecules in the outer epidermal walls of mature wheat stems (triticum aestivum). *Annals of Botany*, 97(6):1091–1094, 2006.

[144] J. S. Lupoi, E. Gjersing, and M. F. Davis. Evaluating lignocellulosic biomass, its derivatives, and downstream products with raman spectroscopy. *Front Bioeng Biotechnol*, 3(50):50, 2015.

[145] H. Schulz, M. Baranska, and R. Baranski. Potential of nir-ft-raman spectroscopy in natural carotenoid analysis. *Biopolymers*, 77(4):212–221, 2005.

[146] A. R. Boyain-Goitia, D. C. Beddows, B. C. Griffiths, and H. H. Telle. Single-pollen analysis by laser-induced breakdown spectroscopy and raman microscopy. *Appl Opt*, 42(30):6119–32, 2003.

[147] Chuji Wang, Yong-Le Pan, Coralyn Hill, and Brandon Redding. Photophoretic trapping-raman spectroscopy for single pollens and fungal spores trapped in air. *Journal of Quantitative Spectroscopy and Radiative Transfer*, 153:4–12, 2015.

[148] A. S. Mondol, M. D. Patel, J. Ruger, C. Stiebing, A. Kleiber, T. Henkel, J. Popp, and I. W. Schie. Application of high-throughput screening raman spectroscopy (hts-rs) for label-free identification and molecular characterization of pollen. *Sensors (Basel)*, 19(20):4428, 2019.

[149] B. G. Pummer, H. Bauer, J. Bernardi, S. Bleicher, and H. Grothe. Suspendable macromolecules are responsible for ice nucleation activity of birch and conifer pollen. *Atmospheric Chemistry and Physics*, 12(5):2541–2550, 2012.

[150] Bernhard G. Pummer, Heidi Bauer, Johannes Bernardi, Bertrand Chazallon, SÃ©bastien Facq, Bernhard Lendl, Karin Whitmore, and Hinrich Grothe. Chemistry and morphology of dried-up pollen suspension residues. *Journal of Raman Spectroscopy*, 44(12):1654–1658, 2013.

[151] Monika Chylinska, Monika Szymanska-Chargot, and Artur Zdunek. Imaging of polysaccharides in the tomato cell wall with raman microspectroscopy. *Plant Methods*, 10(1):14, 2014.

[152] Zsuzsanna Heiner, Ingrid Zeise, Rivka Elbaum, and Janina Kneipp. Insight into plant cell wall chemistry and structure by combination of multiphoton microscopy with raman imaging. *Journal of Biophotonics*, 11(4):e201700164, 2018.

[153] F. Schulte, U. Panne, and J. Kneipp. Molecular changes during pollen germination can be monitored by raman microspectroscopy. *Journal of Biophotonics*, 3(8-9):542–547, 2010.

[154] Virginia Joseph, Manuel Gensler, Stephan Seifert, Ulrich Gernert, JÃ¼rgen P. Rabe, and Janina Kneipp. Nanoscopic properties and application of mix-and-match plasmonic surfaces for microscopic sers. *The Journal of Physical Chemistry C*, 116(12):6859–6865, 2012.

[155] M. Fleischmann, P. J. Hendra, and A. J. McQuillan. Raman spectra of pyridine adsorbed at a silver electrode. *Chemical Physics Letters*, 26(2):163–166, 1974.

[156] K. Kneipp, H. Kneipp, I. Itzkan, R. R. Dasari, and M. S. Feld. Ultrasensitive chemical analysis by raman spectroscopy. *Chem Rev*, 99(10):2957–76, 1999.

[157] J. Kneipp, H. Kneipp, M. McLaughlin, D. Brown, and K. Kneipp. In vivo molecular probing of cellular compartments with gold nanoparticles and nanoaggregates. *Nano Lett*, 6(10):2225–31, 2006.

[158] Janina Kneipp, Harald Kneipp, Burghardt Wittig, and Katrin Kneipp. Novel optical nanosensors for probing and imaging live cells. *Nanomedicine: Nanotechnology, Biology and Medicine*, 6(2):214–226, 2010.

[159] D. Cialla-May, X. S. Zheng, K. Weber, and J. Popp. Recent progress in surface-enhanced raman spectroscopy for biological and biomedical applications: from cells to clinics. *Chemical Society Reviews*, 46(13):3945–3961, 2017.

[160] Mohamed Hassoun, Iwan W Schie, Tatiana Tolstik, Sarmiza E Stanca, Christoph Krafft, and Juergen Popp. Surface-enhanced raman spectroscopy of cell lysates mixed with silver nanoparticles for tumor classification. *Beilstein Journal of Nanotechnology*, 8(1):1183–1190, 2017.

[161] Vesna Zivanovic, Geo Semini, Michael Laue, Daniela Drescher, Toni Aebischer, and Janina Kneipp. Chemical mapping of leishmania infection in live cells by sers microscopy. *Analytical Chemistry*, 90(13):8154–8161, 2018.

[162] A. Matschulat, D. Drescher, and J. Kneipp. Surface-enhanced raman scattering hybrid nanoprobe multiplexing and imaging in biological systems. *ACS Nano*, 4(6):3259–3269, 2010.

[163] V. Zivanovic, S. Seifert, D. Drescher, P. Schrade, S. Werner, P. Guttmann, G. P. Szekeres, S. Bachmann, G. Schneider, C. Arenz, and J. Kneipp. Optical nanosensing of lipid accumulation due to enzyme inhibition in live cells. *ACS Nano*, 13(8):9363–9375, 2019.

[164] Omar E. Rivera-Betancourt, Russell Karls, Benjamin Grosse-Siestrup, Shelly Helms, Frederick Quinn, and Richard A. Dluhy. Identification of mycobacteria based on spectroscopic analyses of mycolic acid profiles. *Analyst*, 138(22):6774–6785, 2013.

[165] Haibo Zhou, Danting Yang, Natalia P. Ivleva, Nicoleta E. Mircescu, Reinhard Niessner, and Christoph Haisch. Sers detection of bacteria in water by in situ coating with ag nanoparticles. *Analytical Chemistry*, 86(3):1525–1533, 2014.

[166] A. Sengupta, M. L. Laucks, and E. J. Davis. Surface-enhanced raman spectroscopy of bacteria and pollen. *Appl Spectrosc*, 59(8):1016–23, 2005.

[167] Alan M. Race, Rory T. Steven, Andrew D. Palmer, Iain B. Styles, and Josephine Bunch. Memory efficient principal component analysis for the dimensionality reduction of large mass spectrometry imaging data sets. *Analytical Chemistry*, 85(6):3071–3078, 2013.

[168] Hiroshi Akima. A new method of interpolation and smooth curve fitting based on local procedures. *J. ACM*, 17(4):589â€“602, 1970.

[169] Andrew N. Krutchinsky and Brian T. Chait. On the nature of the chemical noise in maldi mass spectra. *Journal of the American Society for Mass Spectrometry*, 13(2):129–134, 2002.

[170] Peter Lasch. Spectral pre-processing for biomedical vibrational spectroscopy and microspectroscopic imaging. *Chemometrics and Intelligent Laboratory Systems*, 117:100–114, 2012.

[171] Dong Wei, Shuo Chen, and Quan Liu. Review of fluorescence suppression techniques in raman spectroscopy. *Applied Spectroscopy Reviews*, 50(5):387–406, 2015.

[172] T. Bocklitz, A. Walter, K. Hartmann, P. Rosch, and J. Popp. How to pre-process raman spectra for reliable and stable models? *Anal Chim Acta*, 704(1-2):47–56, 2011.

[173] Nils Kristian Afseth, Vegard Herman Segtnan, and Jens Petter Wold. Raman spectra of biological samples: A study of preprocessing methods. *Applied Spectroscopy*, 60(12):1358–1367, 2006.

[174] A. Savitzky and M. J. E. Golay. Smoothing + differentiation of data by simplified least squares procedures. *Analytical Chemistry*, 36(8):1627–1639, 1964.

[175] yang xi, Yuee Li, zhi Duan, and yang lu. A novel pre-processing algorithm based on the wavelet transform for raman spectrum. *Applied Spectroscopy*, 72:000370281878969, 2018.

[176] Waltraud Kessler. *Datenvorverarbeitung bei Spektren*, book section Datenvorverarbeitung bei Spektren, pages 183–210. Wiley, 2004.

[177] P. H. Eilers. A perfect smoother. *Anal Chem*, 75(14):3631–6, 2003.

[178] P. H. C. Eilers and H. F. M. Boelens. Baseline correction with asymmetric least squares smoothing. *Unpublished*, 2005.

[179] E. T. Whittaker. On a new method of graduation. *Proceedings of the Edinburgh Mathematical Society*, 41:63–75, 1922.

[180] K. Ramser, E. Malinina, and S. Candefjord. Resonance micro-raman investigations of the rat medial preoptic nucleus: Effects of a low-iron diet on the neuroglobin content. *Applied Spectroscopy*, 66(12):1454–1460, 2012.

[181] C. H. Camp, Y. J. Lee, and M. T. Cicerone. Quantitative, comparable coherent anti-stokes raman scattering (cars) spectroscopy: correcting errors in phase retrieval. *Journal of Raman Spectroscopy*, 47(4):408–415, 2016.

[182] Zewei Chen, Zhuoyong Zhang, Ruohua Zhu, Yuhong Xiang, Yuping Yang, and Peter B. Harrington. Application of terahertz time-domain spectroscopy combined with chemometrics to quantitative analysis of imidacloprid in rice samples. *Journal of Quantitative Spectroscopy and Radiative Transfer*, 167:1–9, 2015.

[183] Najla AlMasoud, Yun Xu, Nicoletta Nicolaou, and Royston Goodacre. Optimization of matrix assisted desorption/ionization time of flight mass spectrometry (maldi-tof-ms) for the characterization of bacillus and brevibacillus species. *Analytica Chimica Acta*, 840:49–57, 2014.

[184] N. Nicolaou, Y. Xu, and R. Goodacre. Detection and quantification of bacterial spoilage in milk and pork meat using maldi-tof-ms and multivariate analysis. *Anal Chem*, 84(14):5951–8, 2012.

[185] Loong Chuen Lee and Abdul Aziz Jemain. Predictive modelling of colossal atr-ftir spectral data using pls-da: empirical differences between pls1-da and pls2-da algorithms. *Analyst*, 144(8):2670–2678, 2019.

[186] S. O. Deininger, D. S. Cornett, R. Paape, M. Becker, C. Pineau, S. Rauser, A. Walch, and E. Wolski. Normalization in maldi-tof imaging datasets of proteins: practical considerations. *Anal Bioanal Chem*, 401(1):167–81, 2011.

[187] P. Bassan and P. Gardner. *Scattering in Biomedical Infrared Spectroscopy*. Biomedical Applications of Synchrotron Infrared Microspectroscopy. Royal Soc Chemistry, Cambridge, 2011.

[188] Nils Kristian Afseth and Achim Kohler. Extended multiplicative signal correction in vibrational spectroscopy, a tutorial. *Chemometrics and Intelligent Laboratory Systems*, 117:92–99, 2012.

[189] Sabrina Diehn. *Charakterisierung von Gräserpollen durch multivariate statistische Auswertung verschiedener Spektren*. Master thesis, 2016.

[190] Bruce Dien, Gautam Sarath, Jeffrey Pedersen, Scott Sattler, Han Chen, and Deanna Funnell-Harris. Improved sugar conversion and ethanol yield for forage sorghum (sorghum bicolor l. moench) lines with reduced lignin contents. *BioEnergy Research*, 2, 2009.

[191] Fikadu Biru. *Influenceof Silicon On Sorghum (Sorghum Bicolor L.) Detached leaves SENESCENCE*. Thesis, 2018.

[192] Sabine Holz, Michael Kube, Grzegorz Bartoszewski, Bruno Huettel, and Carmen BÃ¼ttner. Initial studies on cucumber transcriptome analysis under silicon treatment. *Silicon*, 2015.

[193] Santosh Kumar, Yonat Milstein, Yaniv Brami, Michael Elbaum, and Rivka Elbaum. Mechanism of silica deposition in sorghum silica cells. *New Phytologist*, 213(2):DOI: 10.1111/nph.14173, 2016.

[194] B. Zimmermann and A. Kohler. Optimizing savitzky-golay parameters for improving spectral resolution and quantification in infrared spectroscopy. *Appl Spectrosc*, 67(8):892–902, 2013.

[195] P. C. Lee and D. Meisel. Adsorption and surface-enhanced raman of dyes on silver and gold sols. *The Journal of Physical Chemistry*, 86(17):3391–3395, 1982.

[196] K. A. Solhaug. Influence of photoperiod and temperature on dry matter production and chlorophyll content in temperate grasses [also incl. net assimilation rate, nar, long days, short days]. *Norwegian journal of agricultural sciences*, 5(4):365–384, 1991.

[197] R. J. Porra, W. A. Thompson, and P. E. Kriedemann. Determination of accurate extinction coefficients and simultaneous equations for assaying chlorophylls a and b extracted with four different solvents: verification of the concentration of chlorophyll standards by atomic absorption spectroscopy. *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, 975(3):384–394, 1989.

[198] Oshry Markovich, Evyatar Steiner, Stepan Kouril, Petr Tarkowski, Asaph Aharoni, and Rivka Elbaum. Silicon promotes cytokinin biosynthesis and delays senescence in arabidopsis and sorghum. *Plant, Cell and Environment*, 40(7):1189–1196, 2017.

[199] William H. Kruskal and W. Allen Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47(260):583–621, 1952.

[200] Maurice Stevenson Bartlett and Ralph Howard Fowler. Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London. Series A - Mathematical and Physical Sciences*, 160(901):268–282, 1937.

[201] Michael W. Berry, Murray Browne, Amy Nicole Langville, Victor PaÃºl Pauca, and Robert J. Plemmons. Algorithms and applications for approximate nonnegative matrix factorization. *Comput. Stat. Data Anal.*, 52:155–173, 2007.

[202] Thomas Udelhoven, Dieter Naumann, and Jürgen Schmitt. Development of a hierarchical classification system with artificial neural networks and ft-ir spectra for the identification of bacteria. *Applied Spectroscopy*, 54(10):1471–1479, 2000.

[203] Enrico Pigorsch. Spectroscopic characterisation of cationic quaternary ammonium starches. *Starch - Stärke*, 61(3-4):129–138, 2009.

[204] Vishnu Vardhan Pully and Cees Otto. The intensity of the 1602 cmâˆ’1band in human cells is related to mitochondrial activity. *Journal of Raman Spectroscopy*, 40(5):473–475, 2009.

[205] S. Stewart and P. M. Fredericks. Surface-enhanced raman spectroscopy of amino acids adsorbed on an electrochemically prepared silver surface. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 55(7-8):1641–1660, 1999.

[206] Sang Kyu Kim, Myung Soo Kim, and Se Won Suh. Surface-enhanced raman scattering (sers) of aromatic amino acids and their glycyl dipeptides in silver sol. *Journal of Raman Spectroscopy*, 18(3):171–175, 1987.

[207] Peter de B. Harrington, Nancy E. Vieira, Jimmy Espinoza, Jyh Kae Nien, Roberto Romero, and Alfred L. Yergey. Analysis of varianceâ€“principal component analysis: A soft tool for proteomic discovery. *Analytica Chimica Acta*, 544(1-2):118–127, 2005.

[208] Jeroen J. Jansen, Huub C. J. Hoefsloot, Jan van der Greef, Marieke E. Timmerman, Johan A. Westerhuis, and Age K. Smilde. Asca: analysis of multivariate data obtained from an experimental design. *Journal of Chemometrics*, 19(9):469–481, 2005.

[209] A. K. Smilde, J. J. Jansen, H. C. Hoefsloot, R. J. Lamers, J. van der Greef, and M. E. Timmerman. Anova-simultaneous component analysis (asca): a new tool for analyzing designed metabolomics data. *Bioinformatics*, 21(13):3043–8, 2005.

[210] J. De Gelder, K. De Gussem, P. Vandenabeele, and L. Moens. Reference database of raman spectra of biological molecules. *Journal of Raman Spectroscopy*, 38(9):1133–1147, 2007.

[211] F. Tuinstra and J. L. Koenig. Raman spectrum of graphite. *The Journal of Chemical Physics*, 53(3):1126–1130, 1970.

[212] J. Heslop-Harrison and Y. Heslop-Harrison. The growth of the grass pollen tube: 1. characteristics of the polysaccharide particles (p-particles) associated with apical growth. *Protoplasma*, 112(1):71–80, 1982.

[213] Franziska Hanke, Bram J.A. Mooij, Freek Ariese, and Ute Böttger. The evaluation of time-resolved raman spectroscopy for the suppression of background fluorescence from space-relevant samples. *Journal of Raman Spectroscopy*, 50(7):969–982, 2019.

[214] Martin Kögler, Jaakko Itkonen, Tapani Viitala, and Marco G. Casteleijn. Assessment of recombinant protein production in e. coli with time-gated surface enhanced raman spectroscopy (tg-sers). *Scientific Reports*, 10(1):2472, 2020.

[215] L. Mander, M. Li, W. Mio, C. C. Fowlkes, and S. W. Punyasena. Classification of grass pollen through the quantitative analysis of surface ornamentation and texture. *Proc Biol Sci*, 280(1770):20131905, 2013.

[216] G. G. Franchi, B. Piotto, M. Nepi, C. C. Baskin, J. M. Baskin, and E. Pacini. Pollen and seed desiccation tolerance in relation to degree of developmental arrest, dispersal, and survival. *J Exp Bot*, 62(15):5267–81, 2011.

[217] A. R. Lansac, C. Y. Sullivan, B. E. Johnson, and K. W. Lee. Viability and germination of the pollen of sorghum [sorghum bicolor (l.) moench]. *Ann Bot*, 74(1):27–33, 1994.

[218] Frank S. Parker. *Applications of Infrared, Raman, and Resonance Raman Spectroscopy in Biochemistry*. Plenum Press, New York, 1983.

[219] Anna de Juan and RomÃ Tauler. Multivariate curve resolution (mcr) from 2000: Progress in concepts and applications. *Critical Reviews in Analytical Chemistry*, 36(3-4):163–176, 2006.

[220] M. K. Raczkowska, P. Koziol, S. Urbaniak-Wasik, C. Paluszkiewicz, W. M. Kwiatek, and T. P. Wrobel. Influence of denoising on classification results in the context of hyperspectral data: High definition ft-ir imaging. *Anal Chim Acta*, 1085:39–47, 2019.

[221] C. Hughes, A. Henderson, M. Kansiz, K. M. Dorling, M. Jimenez-Hernandez, M. D. Brown, N. W. Clarke, and P. Gardner. Enhanced ftir bench-top imaging of single biological cells. *Analyst*, 140(7):2080–5, 2015.

[222] A. Kohler, C. Kirschner, A. Oust, and H. Martens. Extended multiplicative signal correction as a tool for separation and characterization of physical and chemical information in fourier transform infrared microscopy images of cryo-sections of beef loin. *Appl Spectrosc*, 59(6):707–16, 2005.

[223] E. Döring, J. Schneider, Khidir Hilu, and Martin Röser. Phylogenetic relationships in the aveneae/poeae complex (pooideae, poaceae). *Kew Bulletin*, 62:407–424, 2007.

[224] Anne Blaner, Julia Schneider, and Martin Röser. Phylogenetic relationships in the grass family (poaceae) based on the nuclear single copy locus topoisomerase 6 compared with chloroplast dna. *Systematics and Biodiversity*, 12(1):111–124, 2014.

[225] Andrew R. Korte, Gargey B. Yagnik, Adam D. Feenstra, and Young Jin Lee. *Multiplex MALDI-MS Imaging of Plant Metabolites Using a Hybrid MS System*, pages 49–62. Springer New York, New York, NY, 2015.

[226] Wolfgang Petrich. *From Study Design to Data Analysis*, pages 315–332. 2007.

[227] Esben Jannik Bjerrum, Mads Glahder, and Thomas Skov. Data augmentation of spectral data for convolutional neural network (cnn) based deep chemometrics. *arXiv preprint arXiv:1710.01927*, 2017.

[228] Hong Wang, Yunchao Xie, Dawei Li, Heng Deng, Yunxin Zhao, Ming Xin, and Jian Lin. Rapid identification of x-ray diffraction patterns based on very limited data by interpretable convolutional neural networks. *Journal of Chemical Information and Modeling*, 60(4):2004–2011, 2020.

[229] Kristian Hovde Liland, Achim Kohler, and Volha Shapaval. Hot plsâ€"a framework for hierarchically ordered taxonomic classification by partial least squares. *Chemometrics and Intelligent Laboratory Systems*, 138:41–47, 2014.

[230] Jenny Leopold, Yulia Popkova, Kathrin M. Engel, and Jürgen Schiller. Recent developments of useful maldi matrices for the mass spectrometric characterization of lipids. *Biomolecules*, 8(4):173, 2018.

[231] A. C. M. Veloo, H. Jean-Pierre, U. S. Justesen, T. Morris, E. Urban, I. Wybo, M. Kostrzewa, A. W. Friedrich, T. Morris, H. Shah, H. Jean-Pierre, U. S. Justesen, I. Wybo, E. Nagy, E. Urban, M. Kostrzewa, A. Veloo, and A. W. Friedrich. Validation of maldi-tof ms biotyper database optimized for anaerobic bacteria: The enria project. *Anaerobe*, 54:224–230, 2018.

[232] E. Bille, B. Dauphin, J. Leto, M. E. Bougnoux, J. L. Beretti, A. Lotz, S. Suarez, J. Meyer, O. Join-Lambert, P. Descamps, N. Grall, F. Mory, L. Dubreuil, P. Berche, X. Nassif, and A. Ferroni. Maldi-tof ms andromas strategy for the routine identification of bacteria, mycobacteria, yeasts, aspergillus spp. and positive blood cultures. *Clinical Microbiology and Infection*, 18(11):1117–1125, 2012.

[233] Jennifer Mesureur, Sandrine Arend, BÃ©atrice CelliÃ¨re, Priscillia Courault, Pierre-Jean Cotte-Pattat, Heather Totty, Parampal Deol, Virginie Mick, Victoria Girard, Joanne Touchberry, Vanessa Burrowes, Jean-Philippe Lavigne, David Oâ€™Callaghan, ValÃ©rie Monnin, and Anne Keriel. A maldi-tof ms database with broad genus coverage for species-level identification of brucella. *PLOS Neglected Tropical Diseases*, 12(10):e0006874, 2018.

[234] H. Kajiwara, N. Hinomoto, and T. Gotoh. Mass fingerprint analysis of spider mites (acari) by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry for rapid discrimination. *Rapid Commun Mass Spectrom*, 30(8):1037–42, 2016.

[235] A. Miki, M. Katagi, T. Kamata, K. Zaitsu, M. Tatsuno, T. Nakanishi, H. Tsuchihashi, T. Takubo, and K. Suzuki. Maldi-tof and maldi-fticr imaging mass spectrometry of methamphetamine incorporated into hair. *J Mass Spectrom*, 46(4):411–6, 2011.

[236] Nico Verbeeck, Richard M. Caprioli, and Raf Van de Plas. Unsupervised machine learning for exploratory data analysis in imaging mass spectrometry. *Mass Spectrometry Reviews*, 39(3):245–291, 2020.

[237] Umesh P. Agarwal and Sally A. Ralph. Ft-raman spectroscopy of wood: Identifying contributions of lignin and carbohydrate polymers in the spectrum of black spruce (picea mariana). *Applied Spectroscopy*, 51(11):1648–1655, 1997.

[238] P. N. Perera, M. Schmidt, P. J. Schuck, and P. D. Adams. Blind image analysis for the compositional and structural characterization of plant cell walls. *Anal Chim Acta*, 702(2):172–7, 2011.

[239] U. P. Agarwal. An overview of raman spectroscopy as applied to lignocellulosic materials. *Advances in lignocellulosics characterization. TAPPI Press, Argyropoulos DS (ed)*, Chapter 9:201â€“225, 1999.

[240] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Comput. Surv.*, 31(3):264â€“323, 1999.

[241] P. H. C. Eilers and J. J. Goeman. Enhancing scatterplots with smoothed densities. *Bioinformatics*, 20(5):623–628, 2004.

[242] Ming Hao, Umeshwar Dayal, Ratnesh Sharma, Daniel Keim, and Halldor Janetzko. Visual analytics of large multidimensional data using variable binned scatter plots, 2010.

[243] David W. Scott. On optimal and data-based histograms. *Biometrika*, 66(3):605–610, 1979.

[244] H. J. Bohnert, D. E. Nelson, and R. G. Jensen. Adaptations to environmental stresses. *The Plant Cell*, 7(7):1099–1111, 1995.

[245] A. B. Nicotra, O. K. Atkin, S. P. Bonser, A. M. Davidson, E. J. Finnegan, U. Mathesius, P. Poot, M. D. Purugganan, C. L. Richards, F. Valladares, and M. van Kleunen. Plant phenotypic plasticity in a changing climate. *Trends in Plant Science*, 15(12):684–692, 2010.

[246] Christian Hermans, John P. Hammond, Philip J. White, and Nathalie Verbruggen. How do plants respond to nutrient shortage by biomass allocation? *Trends in Plant Science*, 11(12):610–617, 2006.

[247] A. Bizzini and G. Greub. Matrix-assisted laser desorption ionization time-of-flight mass spectrometry, a revolution in clinical microbial identification. *Clin Microbiol Infect*, 16(11):1614–9, 2010.

[248] Christian J. F. Bertens, Shuo Zhang, Roel J. Erckens, Frank J. H. M. van den Biggelaar, Tos T. J. M. Berendschot, Carroll A. B. Webers, Rudy M. M. A. Nuijts, and Marlies Gijs.

Pipeline for the removal of hardware related artifacts and background noise for raman spectroscopy. *MethodsX*, 7:100883, 2020.

[249] Shuxia Guo, Thomas Bocklitz, and JÃ¼rgen Popp. Optimization of raman-spectrum baseline correction in biological application. *Analyst*, 141(8):2396–2404, 2016.

[250] A. Tfayli, C. Gobinet, V. Vrabie, R. Huez, M. Manfait, and O. Piot. Digital dewaxing of raman signals: discrimination between nevi and melanoma spectra obtained from paraffin-embedded skin biopsies. *Appl Spectrosc*, 63(5):564–70, 2009.

[251] P. Meksiarun, M. Ishigaki, V. A. Huck-Pezzei, C. W. Huck, K. Wongravee, H. Sato, and Y. Ozaki. Comparison of multivariate analysis methods for extracting the paraffin component from the paraffin-embedded cancer tissue spectra for raman imaging. *Sci Rep*, 7(1):44890, 2017.

[252] E. Ly, O. Piot, R. Wolthuis, A. Durlach, P. Bernard, and M. Manfait. Combination of ftir spectral imaging and chemometrics for tumour detection from paraffin-embedded biopsies. *Analyst*, 133(2):197–205, 2008.

[253] Weiwei Tang, Jun Chen, Jie Zhou, Junyue Ge, Ying Zhang, Ping Li, and Bin Li. Quantitative maldi imaging of spatial distributions and dynamic changes of tetrandrine in multiple organs of rats. *Theranostics*, 9(4):932–944, 2019.

[254] Rasmus Bro. Parafac. tutorial and applications. *Chemometrics and Intelligent Laboratory Systems*, 38(2):149–171, 1997.

List of relevant bands in spectra

Table 10.1: List of bands in FTIR spectra of pollen and their tentative assignments.

wavenumber [cm^{-1}]	vibration	tentative assignment
835	ring vibration	sporopollenin[10]
941	C-O-C and C-OH stretch	carbohydrates[10]
963	C-O-C and C-OH stretch	carbohydrates[10]
989	C-O-C and C-OH stretch	carbohydrates[10]
1026	C-O-C and C-OH stretch	carbohydrates[10]
1045	C-O-C and C-OH stretch	carbohydrates, lipids[10]
1066	C-O-C and C-OH stretch	carbohydrates[10]
1079	C-O-C and C-OH stretch	carbohydrates[10]
1166,	ring vibration	sporopollenin[10]
1236	C-O stretch	lipids[5]
1250	amide III	proteins[4]
1331	C-C, C-O skeletal vibration	lipids
1467	CH_2 deformation	lipids[10]
1540	amide II: NH deformation and C-N stretch	proteins[10]
1599	ring vibration	sporopollenin[10]
1624,	C=C	sporopollenin[4]
1649	amide I: C=O stretch	proteins[10]
1669	amide I, C=C stretch	proteins, lipids[80]
1688	amide I, C=O	proteins[4]
1744	C=O stretch	lipids[5, 10]

Table 10.2: List of bands in Raman spectra of pollen and their tentative assignments.

Raman-shift [cm^{-1}]	vibration	tentative assignment
473	C-C-C deformation, skeletal modes	pectin, starch[13, 80, 203]
526	S-S stretch	proteins[13]
938	C-O-C stretch	starch[16, 203]
949	skeletal modes	pectin, starch[13, 80, 203]
1008	C=C ring breathe	proteins, carotenoids[10, 13]
1043		proteins[13]
1082	C-O-C, C-OH deformation	starch[16]
1104	C-OH deformation	Carbohydrates[16]
1126	C-O, C-C stretch	carbohydrates[80]
1161	C-H, C-C stretch	carotenoids[13]
1271	=C-H deformation, amide III	lipids, proteins[31, 80]
1435	CH_2 deformation	lipids[10]
1457	CH_2 deformation	lipids,proteins, carbohydrates[5, 10, 80]
1528		carotenoids[13]
1608		sporopollenin[10]
1654	amide I	proteins[80]
1662	amide I	proteins[10, 80]

Table 10.3: List of bands in SERS spectra of pollen and their tentative assignments.

Raman-shift [cm^{-1}]	vibration	tentative assignment
494		nucleobases[3]
649	ring breathe	nucleobases[3]
735	ring breathe	nucleobases[3]
774		amino acids[205]
802	ring breathe	nucleobases, amino acids[3,205]
921	deformation	nucleobases, amino acids[3,205]
929	C-COO stretch	amino acids[205]
957	deformation	nucleobases[3]
1021	ring breathe	amino acids[3]
1154	NH_3^+ deformation	amino acids[3,205]

Table 10.4: List of bands in Raman spectra of plant tissue cross sections and their tentative assignments.

Raman-shifts [cm^{-1}]	vibration	tentative assignment
896	Skeletal deformation, C-C, C-O stretch	cellulose[6,51,237]
1089	C-O-C stretch, HCC deformation	cellulose[51,237]
1138/1141		suberin[152]
1167	C-C	lignin, carotenoids[13]
1336	OH bend	lignin[51]
1379	HCC, HCO, HOC deformation	cellulose[237]
1519		carotenoids[13]
1598	aromatic ring stretch	lignin[51,237]
1627	C=C stretch	lignin[51,237]
1660	C=C, C=O	lignin[51,237]

List of abbreviations

ANN	artificial neural network
ASCA	ANOVA-simultaneous Component Analysis
AsLS	asymmetric least square
CPC	consensus principal component
CPCA	consensus principal component analysis
EDX	energy dispersive X-ray spectroscopy
EMSC	extended multiplicative scattering correction
FTIR	Fourier-transform infrared spectroscopy
full-CV	leave-one-out cross validation
HCA	hierarchical cluster analysis
HCCA	α-cyano-4-hydroxycinnamic acid
MALDI-TOF MS	matrix-assisted laser desorption/ionization-time-of-flight mass spectrometry
MANOVA	one-way multivariate analysis of variance
MSI	mass spectrometry imaging
NIPALS	nonlinear iterative partial least squares
NMF	non-negative matrix factorization
PC	principal component
PCA	principal component analysis
PLS-DA	partial least squares discriminant analysis
RF	random forest
RMSE	root mean square error
SERS	surface-enhanced Raman scattering
SVD	singular value decomposition
TIC	total ion count

List of Figures

List of Tables

List of publications

Journal articles discussed in this thesis

Matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS) shows adaptation of grass pollen composition, **Diehn, S.**, Zimmermann, B., Bağcıuğlu, M., Seifert, S., Kohler, A., Ohlson, M., Fjellheim, S., Weidner, S., Kneipp, J., *Sci Rep.*;8(1):16591, 2018; doi: 10.1038/s41598-018-34800-1.

Combining Chemical Information From Grass Pollen in Multimodal Characterization, **Diehn S.**, Zimmermann B., Tafintseva V., Seifert S., Bağcıuğlu M., Ohlson M., Weidner, S., Fjellheim, S., Kohler, A., Kneipp, J., *Front Plant Sci.* 10(1788), 2020; doi: 10.3389/f-pls.2019.01788.

Discrimination of grass pollen of different species by FTIR spectroscopy of individual pollen grains, **Diehn S.**, Zimmermann B., Tafintseva V., Bağcıuğlu M., Kohler, A., Ohl-son M., Fjellheim, S., Kneipp, J., *Anal Bioanal Chem.* 2020; doi: 10.1007/s00216-020-02628-2.

Multivariate Raman mapping for phenotypic characterization in plant tissue sections. Liedtke I*, **Diehn S***, Heiner Z*, Seifert S, Obenaus S, Büttner C, Janina Kneipp. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy.* 2021;251:119418.; doi: 10.1016/j.saa.2020.119418.

*Authors contributed equally to the work.

Other journal articles

Multivariate Analysis of MALDI Imaging Mass Spectrometry Data of Mixtures of Single Pollen Grains, Lauer, F., **Diehn, S.**, Seifert, S., Kneipp, J., Sauerland, V., Barahona, C., Weidner, S., *J Am Soc Mass Spectrom*, vol. 29, no. 11, pp. 2237?2247, 2018; doi: 10.1007/s13361-018-2036-5.

Spectroscopic discrimination of sorghum silica phytoliths, Zancajo, V. M. R., **Diehn, S.**, Filiba, N., Goobes, G., Kneipp, J., Elbaum, R., *Front Plant Sci.* 10:1571, 2019; doi: 10.3389/fpls.2019.01571.

Conference contributions

Multivariate imaging for the identification of pollen species in artificial pollen grain mixtures using MALDI-TOF MS,**Diehn, S.**, Lauer, F., Seifert, S., Weidner, S., Kneipp, J., ANAKON 2017, Tübingen, Germany, 3-6. April 2017. (Poster)

Relating information from vibrational spectra to phenotypic variation in plants, **Diehn, S.**, Zeise, I., Heiner, Z., Kohler, A., Kneipp, J., 15^{th} Scandinavian Symposium on Chemometrics, Naantali, Finland, 19-22. June 2017. (Poster)

Multivariate analysis of Raman imaging data to study differences in plant organs, **Diehn, S.**, Zeise, I., Heiner, Z., Kneipp, J., 1^{st} international plant spectroscopy conference, Umea, Sweden, 29-30. August 2017. (Talk)

Hierarchical classification of variations in grass pollen quality using MALDI-TOF MS, **Diehn, S.**, Zimmermann, B., Bağcıuğlu, M., Seifert, S., Kohler, A., Ohlson, M., Fjellheim, S., Weidner, S., Kneipp, J., FTIR Spectroscopy in Micrbiological and Mediacal Diagnostics, Berlin, Germany, 19-20. October 2017. (Poster)

Analysis of plant tissues using vibrational and other spectroscopic methods and multivariate approaches, Zeise, I., Heiner, Z., Joester, M., **Diehn, S.**, Rodriguez, V., Emmerling, F., Elbaum, R., Kneipp, J., FTIR Spectroscopy in Micrbiological and Mediacal Diagnostics, Berlin, Germany, 19-20. October 2017 (Poster, poster prize)

Monitoring of grass pollen quality using MALDI-TOF MS and chemometrics, **Diehn, S.**, Zimmermann, B., Bağcıuğlu, M., Seifert, S., Kohler, A., Ohlson, M., Fjellheim, S., Weidner, S., Kneipp, J., 12. Interdisziplinäres Doktorandenseminar 2018, Berlin, Germany, 25-27. March 2018 (Talk)

Pollen chemistry adapts to stress in plants, **Diehn, S.**, Biru, F., Elbaum R., Kneipp, J., HUJI-HU Workshop, Berlin, Germany, 8-9. October 2018 (Talk)

Raman spectroscopy shows adaption of pollen composition in *Poa alpina*, **Diehn, S.**, Seifert, S., Zimmermann, B., Bağcıuğlu, M., Kohler, A., Ohlson, M., Fjellheim, S., Weidner, S., Kneipp, J., 13. Interdisziplinäres Doktorandenseminar 2019, Berlin, Germany, 18-20. March 2019 (Talk) and 2^{nd} international plant spectroscopy conference, Berlin, Germany, 24-28. March 2019. (Talk)

FTIR spectroscopy of single grass pollen, **Diehn, S.**, Seifert, S., Zimmermann, B., Bağcıuğlu, M., Kohler, A., Ohlson, M., Fjellheim, S., Kneipp, J., Workshop on Machine Learning and Chemometrics in Biospectroscopy, Minsk, Belarus, 18-21. August 2019 (Talk)

Identification of grass pollen species using FTIR microspectroscopy on embedded pollen grains, **Diehn, S.**, Zimmermann, B., Bağcıuğlu, M., Ohlson, M., Fjellheim, S. ,Kohler, A. Kneipp, J., FTIR Spectroscopy in Microbiological and Medical Diagnostics, Berlin, Germany, 10-11. October 2019 (Poster)

FTIR microspectroscopy of organic and inorganic components of plant cells, Zancajo, V., Lindtner, T., **Diehn, S.**, Elbaum, R., Kneipp, J., FTIR Spectroscopy in Microbiological and Medical Diagnostics, Berlin, Germany, 10-11. October 2019 (Poster, poster prize)

Declaration

I declare that I have completed the thesis independently using only the aids and tools specified. I have not applied for a doctor's degree in the doctoral subject elsewhere and do not hold a corresponding doctor's degree. I have taken due note of the Faculty of Mathematics and Natural Sciences PhD Regulations, published in the Official Gazette of Humboldt-Universität zu Berlin no. Nr. 42/2018 on 11.07.2018